НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ УКРАЇНИ
«КИЇВСЬКИЙ ПОЛІТЕХНІЧНИЙ ІНСТИТУТ
імені ІГОРЯ СІКОРСЬКОГО»

Інститут телекомунікаційних систем

Кафедра інформаційно-телекомунікаційних мереж

«На правах рукопису»
УДК 004.021

«До захисту допущено»

Завідувач кафедри

_____ Лариса ГЛОБА

«___»_____2020 р.

Магістерська дисертація

на здобуття ступеня магістра
за освітньо-професійною програмою «Інформаційно-комунікаційні технології»
зі спеціальності 172 «Телекомунікації та радіотехніка»

на тему: «Комплексний метод аналізу та прогнозування лояльності абонентів на основі технології машинного навчання»

Виконала:
студентка ІІ курсу, групи ТІ-91мп
Мороз Анастасія Миколаївна _____

Керівник:
Професор кафедри ІТМ ІТС, професор, д.т.н.
Глоба Лариса Сергіївна _____

Рецензент:
д.т.н., ст.н.с.,
заступник директора з наукової роботи
МАН України
Стрижак Олександр Євгенійович _____

Засвідчую, що у цій магістерській дисертації немає запозичень з праць інших авторів без відповідних посилань.
Студентка _____

Київ – 2020 року

Національний технічний університет України
«Київський політехнічний інститут імені Ігоря Сікорського»
Інститут телекомунікаційних систем
Кафедра Інформаційно-телекомунікаційних мереж

Рівень вищої освіти – другий (магістерський)

Спеціальність – 172 «Телекомунікації та радіотехніка»

Освітньо-професійна програма «Інформаційно-комунікаційні технології»

ЗАТВЕРДЖУЮ

Завідувач кафедри

_____ Лариса ГЛОБА

«___»_____2020 р.

ЗАВДАННЯ
на магістерську дисертацію студентці
Мороз Анастасії Миколаївні

1. Тема дисертації «Комплексний метод аналізу та прогнозування лояльності абонентів на основі технології машинного навчання», науковий керівник дисертації професор кафедри інформаційно-телекомунікаційних мереж ІТС Глоба Лариса Сергіївна, професор, д.т.н., затверджені наказом по університету від «03» листопада 2020 р. № 3208-с

2. Термін подання студентом дисертації 10 грудня 2020р

3. Об’єкт дослідження - процес відтоку клієнтів від оператора мобільного зв’язку.

4. Предмет дослідження - методи статистичного аналізу великих даних, а саме: метод дерева рішень, асоціативних правил та bagging.

5. Перелік завдань, які потрібно розробити:

1. Проаналізувати методи машинного навчання, які використовуються в рамках вирішення проблеми передбачення відтоку абонентів.
2. Провести аналіз вхідних даних, які отримані під час роботи однієї з великих українських операторів мобільного зв’язку.
3. За допомогою вхідних даних навчити систему та провести моделювання обраними методами машинного навчання.

4. Визначати найбільш суттєві фактори, які впливають на рішення абонента припинити користуватися послугами телеком оператора, вибірку абонентів, які найбільш схильні до відтоку та скласти класифікацію абонентів.
5. Запропонувати сценарій бізнес процесу для вирішення проблеми передбачення відтоку абонентів на основі технології машинного навчання.
6. Перевірити працездатність запропонованного сценарію на тестовій вибірці даних.

6. Орієнтовний перелік ілюстративного матеріалу :

1. Тема, актуальність, мета, задачі.
2. Аналіз існуючих рішень.
3. Дані оператору мобільного зв'язку.
4. Модель роботи машинного навчання.
5. Отримані результати передбачення.
6. Публікації.

7. Орієнтовний перелік публікацій:

- UkrMiCo 2018 – Big Data processing for telecom operator system, L Globa, A. Moroz.
- SAIT 2018 – Data Mining and its application, L. Globa, A.Moroz.
- Проблеми телекомунікацій 2019 - Застосування методів Data Mining в телекоммунікаційних системах, Л. Глоба, А. Мороз.
- OSTIS 2019 - Approach to prediction of mobile operators subscribers churn, A. Baria, L. Globa, A. Moroz.

8. Дата видачі завдання 4 вересня 2019 р.

Календарний план

№ з/п	Назва етапів виконання магістерської дисертації	Термін виконання етапів магістерської дисертації	Примітка
1	Аналіз поставленої задачі	4.09.2019 – 3.11.2019	виконано
2	Огляд існуючих методів машинного навчання які застосовуються для рішення задачі відтоку.	5.11.2019 - 30.12.2019	виконано
3	Вибір декількох методів для проведення передбачення відтоку	8.01.2020 - 1.03.2020	виконано
4	Вибір системи для проведення передбачення	2.03.2020 – 30.05.2020	виконано
5	Аналіз та обробка вхідних даних	1.09.2020- 1.10.2020	виконано
6	Встановлення переліку параметрів що оказують найбільший вплив	1.10.2020-1.11.2020	виконано
7	Створення бізнес процесу передбачення відтоку абонентів	1.10.2020-1.11.2020	виконано
8	Підготовка тексту дисертації	1.11.2020 - 7.12.2020	виконано

Студентка Анастасія МОРОЗ

Науковий керівник дисертації Лариса ГЛОБА

РЕФЕРАТ

Випускна робота: 100 сторінок, 11 ілюстрацій, 27 таблиць, 52 джерел;

Мета роботи – підвищення ефективності виявлення факторів та закономірностей із статистичних наборів даних, які впливають на рішення абонента припинити користуватися послугами оператора мобільного зв’язку.

В результаті дослідження були розглянуті такі методи машинного навчання як: асоціативних правил, дерева рішень та bagging. Проаналізовано вхідні данні від одного з найбільших операторів мобільного зв’язку. Запропоновано сценарій бізнес процесу для визначення відтоку абонентів від оператору мобільного зв’язку за допомогою комплексу методів машинного навчання. Отримані фактори, що найбільше впливають на відтік абонентів мобільного оператору та параметри про абонентів, що можуть відмовитись від послуг оператора мобільного зв’язку.

ТЕЛЕКОМ ОПЕРАТОР, ВІДТІК, МАШИННЕ НАВЧАННЯ, БІЗНЕС АНАЛІЗ, ВЕЛИКІ ДАНІ.

ABSTRACT

Final work: 100 pages, 11 illustrations, 27 tables, 52 references.

The purpose of the work is to increase the efficiency of identifying factors and patterns of statistical data sets that affect the subscriber's decision to stop using the services of a mobile operator.

As a result of the study, such machine learning methods as: associative rules, decision trees and bagging were considered. Input from one of the largest mobile operators analyzed. A business model for determining the outflow of subscribers from a mobile operator using a set of machine learning methods is proposed. Regularities and factors influencing the outflow of subscribers of the mobile operator and parameters about the subscribers that may refuse the services of the mobile operator have been obtained.

TELECOM OPERATOR, OUTFLOW, MACHINE LEARNING, BUSINESS ANALYSIS, BIG DATA.

ЗМІСТ

ПЕРЕЛІК СКОРОЧЕНЬ

CART	Classification and Regression Tree
ROC	Receiver operating characteristic
CHAID	Chi-square Automatic Interaction Detector
NoSQL	Not Only SQL
СУБД	Система управління базами даних
TP	True Positives
TN	True Negatives
FN	False Negatives
FP	False Positives
MNP	Mobile Number Portability

ВСТУП

Актуальність теми. Телекомунікаційний сектор став однією із основних галузей промисловості розвинених країн. Щороку провайдери телекомунікаційних послуг зазнають збитків через відтік абонентів. .1 травня 2019 року в Україні стартувала послуга перенесення абонентського номера до іншого оператора (MNP). При такому стані ринку телекомунікаційних послуг, однією із головних проблем компаній-операторів є відтік клієнтів. У цьому конкурентному ринку клієнти надають перевагу високій якості за меншу ціну, у той час як провайдери зосереджені над створенням вигідних пропозицій. Залучення нових клієнтів коштує в кілька разів дорожче, ніж утримання старих. Телеком оператори зацікавлені у наданні якісних послуг великій кількості абонентів. Для цього оператор повинен аналізувати великі дані, щоб вияснити причини незадоволеності абонентів та покащувати якість наданих послуг. Це робить проблему відтоку абонентів особливо актуальною для вивчення.

Технічний прогрес і зростаюча кількість операторів підвищили рівень конкуренції. Компанії наполегливо працюють над тим, щоб вижити на цьому конкурентному ринку залежно від декількох стратегій. Для отримання більших доходів запропоновані основні стратегії: (1) придбати нових клієнтів, (2) продати існуючих клієнтів та (3) збільшити термін утримання клієнтів. Однак порівняння цих стратегій з урахуванням вартості рентабельності інвестицій кожної з них показало, що третя стратегія є найбільш профільною стратегією. Це свідчить про те, що утримання існуючого замовника послуг коштує набагато нижче, ніж залучення нового.

Для застосування третьої стратегії компаніям потрібно впливати на таке явище, відоме як «переміщення клієнта від одного постачальника до іншого». Відтік клієнтів викликає значну стурбованість у секторах послуг з високою конкуренцією. З іншого боку, прогнозування клієнтів, які, ймовірно, покинуть компанію, представлятиме потенційно велике додаткове джерело доходу, якщо це буде зроблено на ранній фазі.

Багато досліджень підтвердили, що методи машинного навчання та аналізу даних дуже ефективні для прогнозування цієї ситуації. Ця методика застосовується за допомогою методів виявлення патернів або причиннонаслідкових зв’язків у поведінці абонентів на основі історичних данихю

Мета і задачі дослідження. Метою роботи є підвищення ефективності виявлення факторів та закономірностей із статистичних наборів даних, які впливають на рішення абонента припинити користуватися послугами оператора мобільного зв’язку.

Для досягнення мети дослідження було поставлено та вирішено такі основні **задачі**:

1. Проаналізувати методи машинного навчання, які використовуються в рамках вирішення проблеми передбачення відтоку абонентів.
2. Провести аналіз вхідних даних, які отримані під час роботи однієї з великих українських операторів мобільного зв’язку.
3. За допомогою вхідних даних навчити систему та провести моделювання обраними методами машинного навчання.
4. Визначати найбільш суттєві фактори, які впливають на рішення абонента припинити користуватися послугами телеком оператора, вибірку абонентів, які найбільш схильні до відтоку та скласти класифікацію абонентів.
5. Запропонувати сценарій бізнес процесу для вирішення проблеми передбачення відтоку абонентів на основі технології машинного навчання.
6. Перевірити працездатність запропонованного сценарію на тестовій вибірці даних.

Об’єкт дослідження –процес відтоку клієнтів від оператора мобільного зв’язку.

Предмет дослідження – методи статистичного аналізу великих даних, а саме: метод дерева рішень, асоціативних правил та bagging.

Наукова новизна одержаних результатів. Науковою новизною роботи є комплексний метод передбачення відтоку абонентів, який поєднує в собі декілька методів машинного навчання та дозволяє визначити закономірності і фактори, які найбільше впливають на відтік для того, щоб ефективно мінімізувати кількість абонентів, які не задоволені послугами телеком оператора.

Найбільш суттєві наукові результати

1. Запропоновано сценарій бізнес процесу, який вирішує задачу передбачення лояльності абонентів.
2. Запропоновано алгоритм проведення аналізу на основі даних оператора, який включає такі етапи як: підготовка даних, навчання, проведення аналізу та прийняття можливого рішення щодо продовження користування послугами абонентом.

Практичне значення одержаних результатів.

1. Проведене математичне моделювання вирішення задачі передбачення відтоку клієнтів на основі запропонованого алгоритму, який визначає рішення абонента щодо задоволеності якості послуг оператора зв'язку методами машинного навчання.
2. Отримано перелік факторів, що мають найбільший вплив на відтік клієнтів.
3. Отримані результати дають можливість телеком операторам вплинути на окремих користувачів та надати кожному з них ті послуги, в яких абонент зацікавлений більше всього для запобігання відтоку клієнтів.

Публікації. Основні результати роботи опубліковані в 1 статті у наукових фахових виданнях, 1 доповідь у працях міжнародних конференцій, 2 в тезах доповідей у працях всеукраїнських конференцій, усього в 4 наукових працях.

РОЗДІЛ 1

BIG DATA І ТЕЛЕКОМУНІКАЦІЙНИЙ СЕКТОР

1.1 Застосування Big Data телеком оператором

На сьогоднішній день телекомунікаційна компанія, яка обслуговує на принципі передоплати 8 мільйонів передплатників послуг мобільного зв'язку, генерує приблизно 30 мільйонів записів про дзвінки (CDR). Тобто, до 11 мільярдів щорічно. Якщо цей же оператор надає ще й послуги мобільного зв'язку на постоплатной основі, а також послуги фіксованого зв'язку, то генерується ще більший обсяг даних[5].

Аналітики компанії IBS «весь світовий об’єм даних» оцінили такими величинами:

2003 г. — 5 ексабайтів даних (1 ЭБ = 1 млрд гігабайтів)

2008 г. — 0,18 зеттабайта (1 ЗБ = 1024 ексабайта)

2015 г. — більше 6,5 зеттабайтів

2020 г. — 40–44 зеттабайта (прогноз)

2025 г. — цей об’єм збільшується ще у 10 разів.

Сьогодні ці дані доступні в формі актуальної інформації в режимі реального часу. Це дозволяє телеком-компаніям в режимі реального часу реагувати на поведінкові зміни в мисленні клієнтів. Це також допомагає реагувати на загрози, пов'язані з конкуренцією на ринку. Телеком - це сектор економіки, в якому Big Data (Великі Дані) виграють битву з традиційними інструментами для проведення бізнес-аналізу.

Великі Дані (Big Data) - це набір надзвичайно великих обсягів даних, які неможливо обробити, використовуючи традиційні інструменти, але які корисні для розвитку бізнесу або суспільству. Ось три критерії для великих даних: величезний обсяг, висока швидкість (постійне і швидке генерування даних з подальшою швидкою обробкою) і велика різноманітність (інформація надходить з різних джерел в структурованої і / або неструктурованою формі). Дані збираються, оброблюються і аналізуються в режимі пакетної обробки або у

вигляді потоків даних в режимі реального часу для отримання корисної інформації низкою зацікавлених сторін[2].

Big Data перетворилися в стратегічний актив телекомунікаційної галузі. Завдяки наявності доступу до великих даних галузь телекомунікацій, по суті, має важливу інформацію, яка дозволяє отримувати капіталізацію, використовуючи ці цінні масиви даних.

Великі Дані корисні телекомунікаційній галузі для утримання абонентів; сегментації клієнтів; оптимізації мережі, планування, а також можливості додаткових / перехресних продажів.

Основні задачі, які стоять перед телеком оператором:

- Поліпшення якості обслуговування клієнтів
- Оптимізація мережі
- Операційна аналітика
- Монетизація знань про абонентів

1.1.1 Поліпшення якості обслуговування клієнтів

Зручність для клієнтів - це ключ до підтримки диференціації ринку і зниження відтоку. Використовуючи аналіз, телекомунікаційні компанії оптимізують і покращують якість обслуговування клієнтів. Також використання Big Data дає провайдерам всебічне уявлення про клієнтів. Це розуміння компанії використовують для мікросегментації абонентської бази і формування цільового і привабливого підходу до обслуговування клієнтів. Дані аналізу також використовуються для складання рекомендацій, а крім того - для прогнозування і прийняття необхідних рішень для запобігання відтоку користувачів[1].

- Цільовий маркетинг.

Провівши аналіз таких характеристик, як моделі поведінки користувачів, дані про виставлені рахунки, запити на підтримку, історія покупок, переваги в плані обслуговування, демографічна інформація, місце розташування і т.д., телекомунікаційні компанії пропонують продукти і послуги на індивідуальній основі. Це також дозволяє компаніям активно представляти потрібну пропозицію в потрібний час - а також в правильному контексті і правильним

клієнтам, що підвищить коефіцієнт конверсії. Наприклад - пропозиція планів поповнення або рекомендації по додатковим пакетам послуг на основі використання Big Data[3].

- Прогнозований відтік клієнтів

Відтік клієнтів впливає на стан телекомунікаційної галузі. Аналіз Великих Даних допомагає об'єднати різні моменти, пов'язані з даними, - якість обслуговування, якість роботи мережі, продуктивність, інформацію про виставлення рахунків передплатникам, відомості про дзвінки в сервісні центри, а також настрої користувачів соціальних мереж. Це створює моделі для прогнозування та прийняття заходів для запобігання відтоку клієнтів.

- Життєвий цикл користувача

Використовуючи дані аналізу, проведеного в режимі реального часу і відображає карту життєвого шляху користувача, телекомунікаційні компанії формують точкові пропозиції і перетворюють зацікавлених осіб в клієнтів. Такі дані, як демографічна інформація про клієнтів, купівельну поведінку, ланцюжок «кліків» мишкою в поєднанні з такими атрибутами, як місце розташування і переваги щодо контенту використовуються для створення нових вигідних пропозицій. Це приносить користь компаніям, так як ті створюють картину взаємодії з конкретними клієнтами на різних етапах життєвого циклу для просування індивідуальних пропозицій і організації рекламних кампаній.

- Підтримка абонента

Використовуючи Big Data, телекомунікаційні компанії створюють інструменти для проведення аналізу, прогнозування та вивчення інформації, щоб заздалегідь виявляти і усувати проблеми. Крім того оператор може надати рішення проблеми, перш ніж та торкнеться клієнта. Провайдери також створили і запустили спеціалізованих ботів, через яких клієнти можуть безпосередньо ставити запитання і отримувати відповіді. Компанії допомагають в активному усуненню проблем, перш ніж ті зроблять негативний вплив на роботу абонентів[10].

1.1.2 Оптимізація мережі

У телекомунікаційному секторі мережу - життєво важливий ресурс. Також важлива ємність мережі. Для моніторингу та управління пропускною спроможністю мережі, побудови прогнозованих моделей пропускної здатності і використання цих моделей при плануванні розширення мережі телекомунікаційний сектор почав використовувати аналітику Big Data[15].

Використовуючи дані, які демонструють кореляцію між використанням мережі і пропускною спроможністю мережі, телекомунікаційні компанії виявляють ділянки з високою завантаженістю, де використання мережі наближається до граничного значення пропускної здатності. Такий аналіз корисний при плануванні розширення ємності мережі.

У регіонах, де спостерігається наявність надлишкової пропускної здатності мережі, телекомунікаційні компанії організують спеціальні кампанії або рекламні акції, спрямовані на збільшення використання ресурсів мережі. На основі аналізу даних і трафіку, одержуваних в режимі реального часу, можуть бути розроблені моделі прогнозування пропускної здатності. На основі аналізу даних, зібраних шляхом порівняння фактичного і прогнозованого трафіку, провайдери можуть планувати додавання в мережу додаткової ємності в разі збоїв. Ці дані також можуть допомогти при виявленні зони скидання виклику і прогнозуванні відповідного положення веж стільникового зв'язку[12].

Телекомунікаційним компаніям необхідно планувати інвестиції на основі ряду параметрів. Таких як майбутні потреби в підключенні, стратегічні цілі, прогнозована рентабельність інвестицій, прогнозований трафік, якість обслуговування клієнтів. Ефективне поєднання даних про трафік мережі, показниках якості обслуговування клієнтів, потенційного прибутку і місцезнаходження в поєднанні з даними про цінності клієнта забезпечують найбільш ефективне використання інвестицій.

Раніше сектор телекомунікацій залежав від історичних даних для управління мережею. Тепер компанії, що працюють в телеком-секторі, почали використовувати Big Data, а також інструменти проведення аналізу для побудови

гарячих карт використання ємності в реальному часі, які відстежують якість взаємодії з користувачем і відправляють оповіщення в разі перевантаження мережі або потенційних збоїв. Аналіз Великих Даних допомагає постійно відстежувати мережеву активність і прогнозувати майбутній попит[14]. Крім того, використовуючи дані, одержувані в реальному часі з веж стільникового зв'язку, інженери здатні відстежити зниження якості послуг, що надаються в певному місці.

1.1.3 Операційна аналітика

Телекомунікаційний сектор використовує аналіз Big Data для операцій, пов'язаних з управлінням компаніями[25]. Зокрема, для таких операцій, як мінімізація відтоку абонентів, управління мережею і кібербезпека, а також для вирішення проблем клієнтів і зниження ризиків, що виникають під час користування послугою.

Для запобігання витоку доходів і шахрайства в цій сфері телекомунікаційні компанії також використовують рішення на основі Великих Даних. Ці рішення допомагають аналізувати як структуровані, так і неструктуровані дані. Цей аналіз допомагає компаніям краще розуміти поведінку клієнтів.

Телекомунікаційна галузь приділяє увагу мережевої безпеки. Для побудови мереж використовуються оптичні волокна, і в цьому випадку виникають проблеми, пов'язані з витоком інформації[35]. Дані, що відносяться до цих небезпек, аналізуються оператором в режимі реального часу, що знижує ризики, виявляє інциденти і дозволяє реагувати на порушення.

1.1.4 Монетизація знань про абонентів

Телекомунікаційні компанії отримують доступ до інформації про місцезнаходження абонента, використання мережею і пристроями, про переваги користувача і т.д. Ця інформація використовується для створення статистики, яка важлива для інших підприємств[4].

Аналіз інформації. Телекомунікаційний сектор почав надавати інформацію, отриману в результаті аналізу даних, як послуги для інших секторів економіки. Для такого аналізу створено багато додатків і різних сценаріїв.

Аналіз IoT / M2M. Телекомунікаційні компанії почали надавати комплексні рішення категорії M2M (machine-to-machine або міжмашинної взаємодії). З огляду на постійне зростання в мережі кількості пристроїв категорії Інтернет речей (IoT), мережева аналітика трафіку IoT-датчиків стає наступною областю проведення досліджень[7]. Тепер телеком-оператори отримали можливість додавання до потоковим даними геолокаційні і геопросторові елементи, в кінцевому підсумку надаючи цінну інформацію для вертикалей підприємства.

Телеком-оператори практично перманентно взаємодіють зі своїми абонентами через смартфони, які стали нерозривною частиною життя будь-якої людини. Саме тому телеком мають найбільш повний профіль клієнта, на основі якого можна створювати прогностичні і рекомендаційні сервіси. Дані сервіси можна вигідно монетизувати, укладаючи співпрацю з банками, рітейлом і навіть державою[9].

Для монетизації компанії можуть використовувати геолокаційні дані, які відіграють надзвичайно важливу роль в транспорті: управління пасажиропотоком, прогнозування попиту на квитки, оптимізація маршрутів і розкладу транспортних засобів.

Ці завдання можуть бути вирішені за допомогою машинного навчання на основі даних про пересування абонентів.

1.2 Принципи роботи з великими даними

Оскільки дані зберігаються на кластері, для роботи з ними потрібна особлива інфраструктура. Найпопулярніша екосистема - це Hadoop. У ній може працювати дуже багато різних систем: спеціальних бібліотек, планувальників, інструментів для машинного навчання і багато чого іншого [17]. Але в першу чергу ця система потрібна, щоб аналізувати великі обсяги даних за рахунок розподілених обчислень.

Наприклад, ми шукаємо найпопулярніший твіт серед даних розбитих на тисячі серверів. На одному сервері ми б просто зробили таблицю і все [28]. Тут

ми можемо притягти всі дані до себе і перерахувати. Але це не правильно, тому що дуже довго.

Тому є Hadoop з парадигмами Map Reduce і фреймворком Spark. Замість того, щоб тягнути дані до себе, вони відправляють до цих даних ділянки програми. Робота йде паралельно, в тисячу потоків. Потім виходить вибірка з тисячі серверів на основі якої можна вибрати найпопулярніший твіт [36].

Map Reduce старіша парадигма, Spark - новіше. З його допомогою дістають дані з кластерів, і в ньому ж будують моделі машинного навчання.

Спираючись на визначення Big Data, можна сформулювати основні принципи роботи з такими даними [54]:

Горизонтальна масштабованість. Оскільки немає обмежень на об'єм інформації, система обробки і зберігання великих даних повинна бути розширюваної: зі збільшенням обсягу даних необхідно пропорційно покращувати апаратну конфігурацію системи.

Відмовостійкість. Можливі виходи з ладу обладнання не повинні робити істотного впливу на методи роботи з великими даними.

Локальність даних. При використанні великих розподілених систем потрібно відповідно безліч обчислювальних машин. І якщо фізично дані розташовані на одному сервері, а обробка виконується на іншому, то це призводить до збільшення витрат, що перевищують часом витрати на саму обробку даних. Тому одним з найважливіших принципів проектування Big Data-рішень є принцип локальності даних - по можливості обробляємо дані на тій же машині, на якій їх зберігаємо.

1.3 Методи аналізу великих даних

Сьогодні в багатьох галузях впроваджують машинне навчання для автоматизації бізнес-процесів і модернізації економічної сфери. Концепція передбачає навчання і керування штучним інтелектом (ШІ) за допомогою спеціальних алгоритмів. Вони вчать систему на основі відкритих даних або отриманого досвіду [22]. Згодом такий додаток здатне прогнозувати розвиток подій без явного програмування людиною і годин витрачених на написання коду.

Наприклад, за допомогою машинного навчання можна створити алгоритм технічного аналізу акцій і передбачуваних цін на них. Використовуючи регресійний і прогнозний аналізи, статистичне моделювання та аналізу дій, експерти створюють програми, які розраховують час вигідних покупок на фондовому ринку [11]. Вони аналізують відкриті дані з бірж і пропонують найбільш вірогідний розвиток подій.

При роботі з Великими даними машинне навчання виконує подібну функцію: спеціальні програми аналізують значні обсяги інформації без втручання людини. Все, що потрібно від оператора «навчити» алгоритм відбирати корисні дані, які потрібні компанії для оптимізації процесів. Завдяки цьому аналітики складають звіти за кілька кліків миші, вивільняючи свій час і ресурси для більш продуктивних завдань: обробки результатів і пошук найбільш ефективних стратегій [43].

У динамічному світі, де очікування клієнтів все вище, а людські ресурси все цінніше, машинне навчання і наука про дані відіграють вирішальну роль у розвитку компанії. Цифрова технологізація робочого процесу життєво необхідна для збереження лідируючих позицій в конкурентному середовищі [50].

Також для обробки великих даних застосовуэться MapReduce - це модель розподіленої обробки даних, запропонована компанією Google для обробки великих обсягів даних на комп'ютерних кластерах.

MapReduce передбачає, що дані організовані у вигляді деяких записів. Обробка даних відбувається в 3 стадії [31]:

Стадія Map. На цій стадії дані предобрабативаются за допомогою функції map (), яку визначає користувач. Робота цієї стадії полягає в передобробці і фільтрації даних. Робота дуже схожа на операцію map в функціональних мовах програмування - призначена для користувача функція застосовується до кожної вхідного відрізку. Функція map () застосована до однієї вхідний записи і видає безліч пар ключ-значеннь. Безліч - тобто може видати тільки один запис, може не видати нічого, а може видати кілька пар ключ-значення. Що буде знаходиться в ключі і в значенні - вирішувати користувачу, але ключ - дуже важлива річ, так

як дані з одним ключем в майбутньому потраплять в один екземпляр функції reduce.

Стадія Shuffle. Проходить непомітно для користувача. У цій стадії висновок функції map «розбирається по кошиках» - кожна корзина відповідає одному ключу виведення стадії map. Надалі ці кошики послужать входом для reduce.

Стадія Reduce. Кожный «кошик» із значеннями, сформований на стадії shuffle, потрапляє на вхід функції reduce (). Функція reduce задається користувачем і обчислює фінальний результат для окремої «кошика». Безліч всіх значень, повернутих функцією reduce (), є фінальним результатом MapReduce-завдання [24].

1.4 Тенденція розвитку великих даних

На сьогоднішний день існують чотири важливих аспекти застосування великих даних: самі дані, аналітика, люди, інструменти. Можна виділити структуровані і неструктуровані дані, і в тому і в іншому вигляді даних можна виділити дані, згенеровані людиною і згенеровані машиною (комп'ютерами, датчиками і т.д.) [38].

Існує цілий ряд проблем в області великих даних, зокрема [45]:

1. Питання якості неструктурованих даних. Можна зіткнутися з неабиякою часткою фальсифікованого контенту в інтернеті. Наприклад, є люди і програми штучного інтелекту для написання відгуків як позитивних, так і негативних.

2. Big Data - це не тільки неструктурована інформація. Так телеком компанії, крім завдань аналізу звернень абнентів, виявлення факторів, що впливають на відтік, переходить до збирання й опрацюваннявеликих обсягів данних, що генеруються телекомунікаційно. системою, які потребують застосування технологій Big Data.

3. Неправильно ставити за мету збір якомога більшої кількості даних, головне - правильно ставити завдання і для їх вирішення шукати правильні дані. Таким чином, необхідно йти не від даних, а від вирішуваних завдань. Збір даних

заради самих даних, захопленість фахівців новими технологіями заради самих технологій у відриві від реальної практики не вирішує поставлених задач.

Висновки

В розділі розглянуто, що телекомунікаційні оператори працюють з надвеликими даними. Великі обсяги інформації генеруються під час роботи телекомунікаційної мережі і містять в собі різнорідну та неструктуровану інформацію.

Телекомунікаційні компанії стикаються з проблемою обробки великих даних, які зумовлені слабкою структурованістю, недостатньою систематизованістю, різнорідністю та слабкозв’язаністю інформації. Для того, щоб ефективно обробляти великі дані потрібно удосконалювати методи аналітичної обробки інформації.

Для удосконалення існуючих систем аналітичної обробки інформації в роботі пропонується застосовувати методи машинного навчання.

РОЗДІЛ 2

АНАЛІЗ ПОКРАЩЕННЯ ОБЧИСЛЮВАЛЬНИХ ПРОЦЕСІВ ДЛЯ РІШЕННЯ ЗАВДАННЯ ПІДВИЩЕННЯ ЕФЕКТИВНОСТІ ДІЯЛЬНОСТІ ТЕЛЕКОМ ОПЕРАТОРА

2.1 Бізнес аналіз

Бізнес-аналіз - це діяльність, яка дозволяє впроваджувати зміни в компанії шляхом визначення потреб і рекомендації рішень, які забезпечують цінність для зацікавлених осіб.

Такий аналіз дозволяє компанії визначити потреби і обґрунтувати зміни, а також розробити і описати рішення, які можуть принести користь. Бізнес-аналіз проводиться в різних ініціативах в рамках компанії. Це можуть бути стратегічні, тактичні, або оперативні ініціативи[18].

Аналіз вимог може здійснюється в межах проекту або в ході розвитку компанії. Його застосовують, щоб зрозуміти поточний стан, визначити майбутній стан, а також визначити дії, які необхідні для переходу від поточного стану до майбутнього.

Існують різні точки зору на бізнес-аналіз: гнучкі методології, інтелектуальний аналіз даних, інформаційні технології, архітектура бізнесу і управління бізнес-процесами. Будь-яку точку зору можна розглядати як призму, через яку бізнес-аналітик розглядає свою робочу діяльність, перебуваючи при цьому в певному контексті[29].

Бізнес-аналіз включає в себе розуміння того, як організації діють для досягнення своїх цілей, і визначення можливостей, потрібних організації для надання продуктів і послуг зовнішнім зацікавленим сторонам. Він включає в себе визначення цілей організації, як ці цілі співвідносяться із завданнями, визначення курсу дій, яких заходів організація повинна вжити для досягнення цих цілей і завдань, а також визначення способу взаємодії різних підрозділів організації та зацікавлених сторін в рамках і поза цієї організації.

Бізнес-аналіз може бути виконаний для розуміння поточного стану організації або служити в якості основи для подальшої ідентифікації потреб бізнесу. Однак, в більшості випадків бізнес-аналіз виконується для визначення і перевірки рішень, які відповідають потребам бізнесу, цілям або завданням[13].

2.2 Ключові концепції бізнес аналізу

Ухвалення рішення неможливо безвідносно до конкретної ситуації. Однією зі складових контексту є особливості, індивідуальність організації та зацікавлених осіб (в т.ч. переконання, культура, стереотипи, стандарти компанії та ін.). Бізнес-аналітик при виборі способів роботи, конкретних методик, розробці рішень спирається не тільки на методології, а й специфіку компанії і особистий досвід. Наприклад, в одних компаніях процес бізнес-аналізу може бути формалізований, стандартизований, підписаний ключовими зацікавленими особами (підвищуючи зобов'язання всіх сторін), а в інших носити менш формальний характер. Підходи до планування варіюються від прогнозних (зводять до мінімуму невизначеність і максимізує контроль) до адаптивних (розрахованих на короткі ітерації). Вибір того чи іншого підходу залежить, як від можливості визначити майбутнє заздалегідь, так і відносини зацікавлених осіб до невизначеності і контролю [30].

Центральна концептуальна модель по бізнес-аналізу (BACCM) - це концептуальний фреймворк для бізнес-аналізу. Модель включає в себе пояснення що таке бізнес-аналіз і що модель означає для тих, хто виконує завдання по бізнес-аналізу, незалежно від поглядів (ракурсів) на бізнес-аналіз, галузі, методології або рівня управління в організації. Модель складається з шести термінів, які мають загальне значення для всіх бізнес-аналітиків і допомагає їм обговорювати бізнес-аналіз і його взаємини з загальноприйнятою термінологією. Кожен з цих термінів є частиною центральної концепції.

Шість ключових концепцій в моделі BACCM:

- Change (Зміна),
- Need (Потреба),
- Solution (Рішення),

- Stakeholder (Зацікавлена сторона),
- Value (Цінність), і
- Context (Контекст).

Кожна ключова концепція - це ідея, яка містить основне значення для практики бізнес-аналізу, а також всі концепції рівні і необхідні. Кожна центральна концепція визначається іншими п'ятьма ключовими концепціями і не можуть бути повністю зрозумілими доти, поки не розкриті всі ключові концепції. Не існує єдиної концепції, яка містить в собі велику важливість або більшої значущості в порівнянні з будь-якою іншою концепцією. Ці концепції грають важливу роль в розумінні типу інформації, яка виявляється, над якою проводиться аналіз або якою управляють в рамках завдань по бізнес-аналізу [48].

Ключові концепції можуть бути використані бізнес-аналітиками для того, щоб розглянути якість і повноту виконаної роботи. У кожній галузі знань є приклади того, як ключові концепції можуть бути використані і / або застосовуватися в ході виконуваних завдань в рамках галузей знань.

Для бізнес аналізу використовується наступна схема класифікації даних, яка описує вимоги [53]:

- Бізнес-вимоги: висловлювання цілей, завдань і результатів, які описують чому було ініційовано зміна. Вони можуть застосовуватися для всього підприємства, бізнес-області або для конкретної ініціативи.
- Вимоги зацікавлених сторін: описують потреби зацікавлених сторін, які повинні бути виконані для задоволення бізнес-вимог. Вони можуть служити в якості моста між бізнес-вимогами та вимогами до Рішення.
- Вимоги до Рішення: описують можливості і якості рішення, які задовольняють вимоги зацікавлених сторін. Вони забезпечують відповідний рівень деталізації, що дозволяє вести розробку і впровадження Рішення. Вимоги до Рішення можна розділити на дві підкатегорії:

- Функціональні вимоги: описують можливості, які Рішення повинно мати в термінах поведінки та інформації, якою Рішення буде управляти;
- Нефункціональні вимоги або вимоги до якості обслуговування: не належать безпосередньо до поведінки функціональності Рішення, а скоріше описують умови, при яких Рішення повинно залишатися ефективним, або якості, яким Рішення повинно задовольняти.
- Перехідні вимоги: описують можливості, які Рішення повинно мати і умови, яким Рішення повинно задовольняти, щоб забезпечити перехід від поточного стану до майбутнього стану, але які будуть не потрібні після того, як зміни будуть здійснені. Ці вимоги відрізняються від інших типів вимог, оскільки вони носять тимчасовий характер. Перехідні вимоги ставляться до таких тем як - перетворення даних, навчання і забезпечення безперервності бізнесу [49].

2.2.1 Класифікація вимог

Виявлення, аналіз, затвердження і управління вимогами неодноразово визнавалися в якості ключових заходів по бізнес-аналізу. Тим не менш, важливо визнати, що бізнес-аналітики також несуть відповідальність за визначення дизайну на певному рівні в ініціативі. Рівень відповідальності за дизайн змінюється в залежності від перспективи (ракурсу), з якими працює бізнес-аналітик [40].

Вимоги сфокусовані на потребах, дизайн сфокусований на вирішенні. Відмінності між вимогами і дизайном не завжди очевидні. Ті ж самі методи використовуються для виявлення, моделювання та аналізу. Вимоги підводять до дизайну, який в свою чергу може зажадати дослідження і аналіз додаткових вимог. Відмінності дуже незначні.

Класифікація вимог і дизайну може стати менш значущою у міру того, як бізнес-аналітик просувається в розумінні потреби і подальшого її задоволення.

Трасування вимог або вказівка і моделювання вимог можуть відноситись до самих вимог, але увагу також слід приділити й дизайну [37].

Бізнес-аналіз може бути складним і рекурсивним. Вимога (або набір вимог) можуть бути використані для визначення дизайну. Дизайн може використовуватися для виявлення додаткових вимог, які використовуються для визначення детальних дизайнів. Бізнес-аналітик може передавати вимоги і дизайни іншим зацікавленим сторонам, які можуть докладніше опрацювати дизайни. Будь то бізнес-аналітик або інша роль, яка завершує розробку дизайнів, бізнес-аналітик часто розглядає остаточні дизайни для того, щоб переконатися що вони відповідають вимогам. У наступній таблиці наведено деякі основні приклади того, як інформація може розглядатися в якості вимоги або дизайну.

Для управління змінами вимог сьогодні найчастіше застосовується підхід, сформульований американським програмним інженером і IT-консультантом Карлом Вігерсом. Основний зміст підходу К. Вігерса сформульовано в його книзі «Розробка вимог до програмного забезпечення» [30].

При створенні ІС виділяються два види вимог: функціональні і нефункціональні. Вігсрс виділяє три види функціональних вимог [8]:

- бізнес-вимоги. Даний вид вимог формується замовником ІС і грунтується, перш за все, на цілях створення замовляється продукту. Бізнес-вимоги визначають, які переваги має отримати замовник при отриманні готового продукту, а також які проблеми або завдання будуть вирішені в результаті його застосування. В результаті формулювання бізнес-вимог окреслюються межі створюваної ІС, а також створюється загальний образ проекту. Наприклад, на рівні бізнес- вимог можуть бути сформульовані вимоги до підтримки бізнес-процесів. Якщо в якості проекту виступає CRM-система, то в якості бізнес-вимог до неї будуть виділені комунікаційні, звітні та управлінські процеси;
- призначені для користувача вимоги - це завдання, які вирішуватиме ІС для підтримки користувачів. Функціональні вимоги даного рівня подаються у вигляді сценаріїв (user journey), алгоритмів і таблиць «подія - відгук».

Також формування призначених для користувача вимог може вестися на основі ключових ролей, які будуть іспользеваться для роботи ИС. Можливості кожної ролі, будь то «Клієнт», «Інвестор», «Партнер» та ін., Будуть відрізнятися в подальшому;

- функціональні вимоги - це основні вимоги по функціональності ІС, які далі детально описуються у вигляді технічного завдання і передаються на реалізацію розробникам.

Але, як показала практика, одних тільки функціональних вимог до ІС недостатньо, оскільки створена лише на їх основі ІС не задовольнятиме всім вимогам бізнесу. У зв'язку з цим К. Вігерс виділяє три види не функціональних вимог до ІС [20]:

- бізнес-правила, що включають вимоги регуляторів (наприклад, екологічні нормативи чи норми безпеки), промислові стандарти, корпоративні стандарти та інші обмеження, які неминуче накладаються зовнішнім середовищем або політикою компанії;
- атрибути якості, які не належать до функціональності системи, однак є обов'язковою умовою для ефективного застосування створюваної ІС в подальшому. Як вимог виду «Атрибути якості» може виступати можливість інтеграції з іншими ІС, інтероперабельність, підтримка програмних продуктів і т.п .;
- обмеження, до яких, як правило, відносяться вимушені технічні або ресурсні обмеження (рівні продуктивності, технічні протоколи та ін.).

2.2.2 Задачі бізнес аналітика

Бізнес-аналітик - це будь-яка особа, яка виконує завдання бізнес-аналізу незалежно від своєї посади або організаційної ролі. Бізнес-аналітик відповідає за виявлення, узагальнення та аналіз інформації з різних джерел в рамках компанії, в тому числі: інструментів, процесів, документації, а також зацікавлених осіб[16].

Бізнес-аналітик відповідає за виявлення реальних потреб зацікавлених осіб (що часто включає в себе аналіз і прояснення які висловлюються побажань) для того, щоб визначити основні завдання та виявити мотиви.

Бізнес-аналітики беруть активну участь в тому, щоб спроектоване і реалізоване рішення співвідносилося з потребами зацікавлених осіб[27]. Зазвичай діяльність бізнес-аналітиків включає в себе:

- осмислення проблем і завдань компанії,
- аналіз потреб і рішень,
- розробка стратегій,
- впровадження змін, і
- допомога у взаємодії зацікавлених осіб.

Бізнес-аналітика також можуть називати:

- бізнес-архітектором,
- аналітиком бізнес-систем,
- аналітиком даних,
- аналітиком підприємств,
- консультантом з питань управління,
- аналітиком процесів,
- менеджером по продукції,
- власником продукту,
- фахівцем з вимогами,
- системним аналітиком.

Коли бізнес-аналітиків не було, вимоги міг готувати сам замовник. Але не завжди це виходило швидко і якісно: у представників великих компаній теж не вистачало ресурсів, часу і знань, щоб створювати документи для програмістів. Часто через погано підготовлені документи терміни релізів зривалися, а це не подобалося ні замовникам, ні виконавцям[23]. З'явилися бізнес-аналітики, і завдяки їм процес збору вимог став швидше і ефективніше.

Від бізнес-аналітиків очікують, що вони ознайомляться з процесами компанії, виявлять, розроблять, інвентаризують та узгодять бізнес-вимоги. Системному аналітику потрібно проаналізувати отримані бізнес-вимоги, уточнити їх, і з огляду на особливості майбутньої системи, розробити детальні функціональні і нефункціональні вимоги. Крім цього, системному аналітику потрібно знати, як прописати вимоги бізнесу так, щоб їх зрозуміли IT-фахівці. Системні аналітики проектують моделі даних, описують протоколи взаємодії між системами[34].

2.2.3 Область знань з бізнес аналізу

Галузі знань представляють собою конкретну експертизу по бізнес-аналізу, яка охоплює кілька завдань. Існує шість областей знань: планування та контроль бізнес аналізу, обстеження і взаємодія, управління життєвим циклом вимог, стратегічний аналіз, аналіз вимог і визначення рішень та оцінка рішень[19].

Планування та контроль бізнес-аналізу: опис завдань, які бізнес-аналітики виконують, щоб організувати роботу і скоординувати зусилля бізнес-аналітиків і зацікавлених осіб. Результати виконання цих завдань використовуються в якості ключових вхідних даних і керівних принципів (рекомендації) для всіх інших завдань керівництва.

Обстеження і взаємодія: описує завдання, які бізнес-аналітики виконують, щоб підготувати і провести обстеження діяльності та затвердити отримані результати[46]. У тому числі описується взаємодія із зацікавленими особами в усіх напрямках діяльності після того, як зібрана інформація для аналізу.

Управління життєвим циклом вимог: описує завдання, які бізнес аналітики виконують для того, щоб управляти і підтримувати вимоги і дані, необхідні для проектування, на всіх етапах життєвого циклу. Ці завдання описують встановлення конструктивних взаємозв'язків між вимогами і дизайном, а також дозволяють оцінювати, аналізувати і приходити до єдиної думки з пропонованими змінами у вимогах і дизайні.

Стратегічний аналіз: описує аналітичну роботу по взаємодії із зацікавленими особами, які мають здійснюватися з метою виявлення стратегічних або тактичних бізнес-потреб, а також привести у відповідність результуючу стратегію з високорівневими і низькорівневими стратегіями.

Аналіз вимог і визначення рішень: описує завдання, які виконують бізнес-аналітики, щоб:

- структурувати і організувати вимоги, виявлені під час обстеження,
- описати їх,
- побудувати модель,
- спроектувати,
- валідувати (перевірити) і верифікувати (затвердити) інформацію,
- визначити варіанти рішень, які відповідають потребам бізнесу, і
- оцінити потенційну цінність, яку може дати кожен варіант рішення.

Ця галузь знань охоплює інкрементальні і ітераційні діяльності: від початкової концепції і дослідження потреб до перетворення цих потреб в приватне рекомендований рішення.

Оцінка рішень: описує завдання, які бізнес-аналітики виконують, щоб оцінити ефективність роботи і цінність рішень, пропонованих компанії-замовнику, а також рекомендувати усунення перешкод або обмежень, які заважають використанню всіх переваг рішення[26].

Всі області знань включають в себе візуальне уявлення вхідних і вихідних даних. Наступна діаграма показує співвідношення між цими галузями знань.

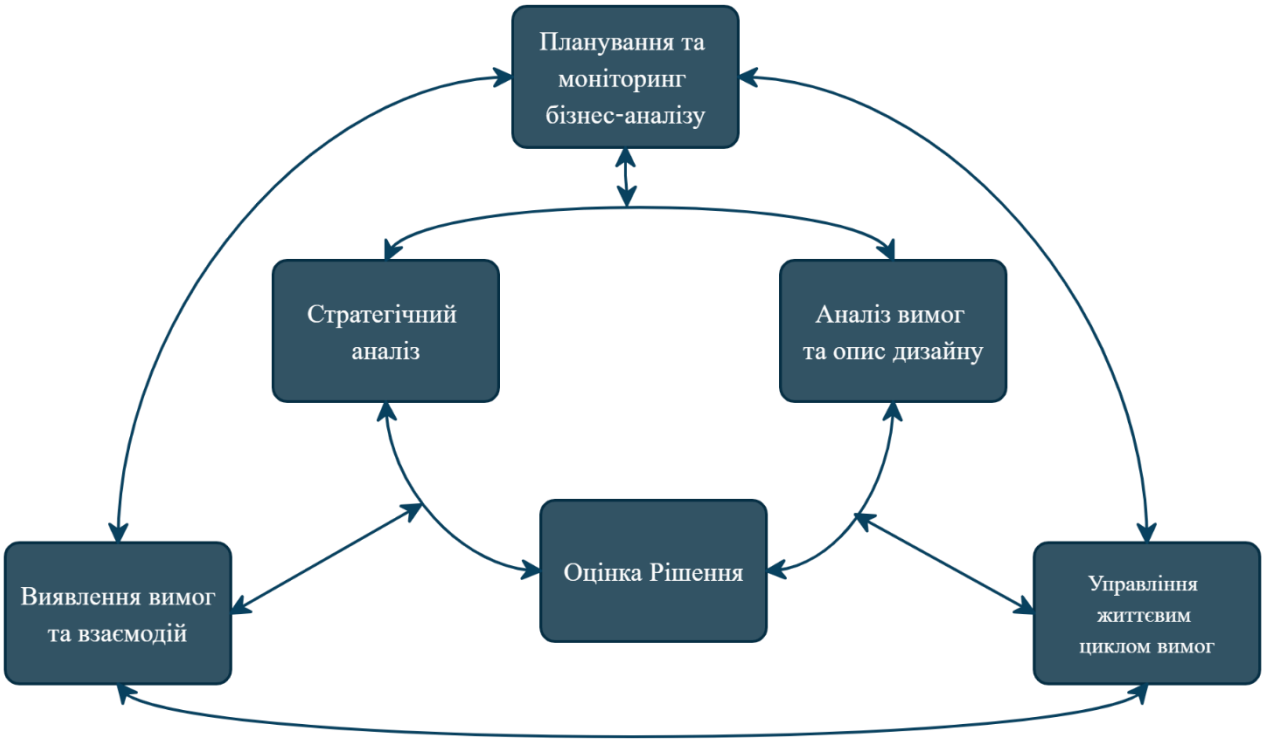

Рис. 2.1 Співвідношення галузі знань

2.3 Задачі бізнес аналізу

Кожна галузь знань описує завдання, які виконуються бізнес-аналітиками для досягнення мети цієї галузі знань[42]. Кожне завдання представлене в наступному форматі:

- Призначення завдання
- Опис завдання
- Вхідні дані
- Діаграма вхідних / вихідних даних завдання
- Вимоги вхідних даних для задачі
- Елементи завдання
- Методи виконання завдання
- Зацікавлені особи
- Вихідні дані завдання

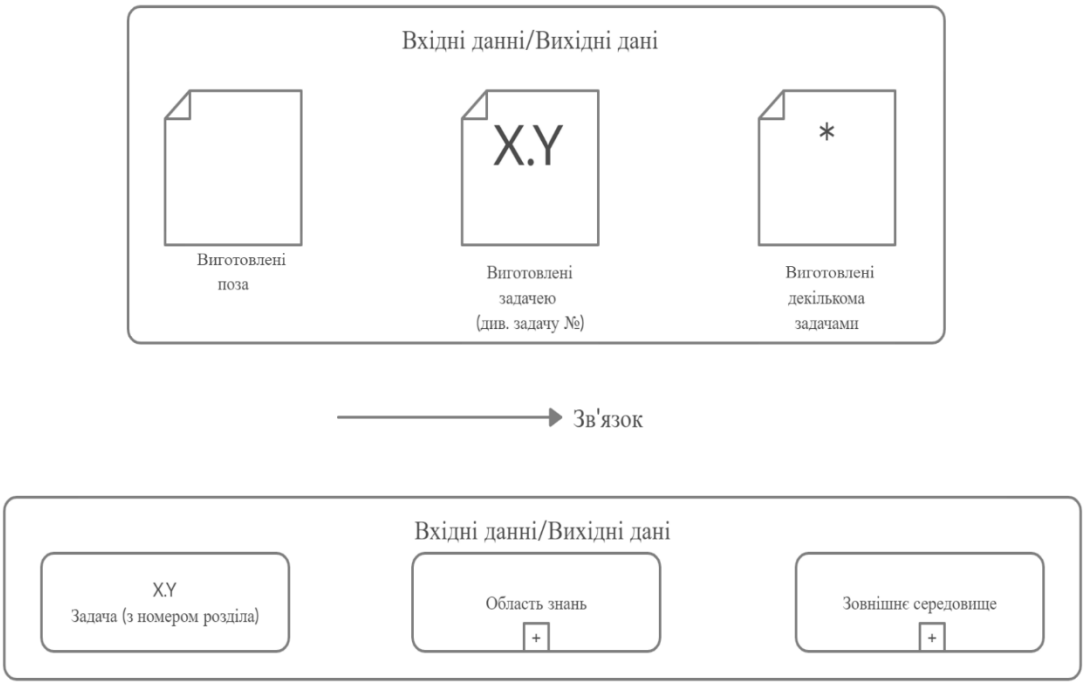

Рис. 2.2 Діаграма вхідних і вихідних даних завдання

Кожне завдання має призначення. Призначення - це короткий опис причини для виконання цього завдання бізнес-аналітиком і її цінності, яка створюється за допомогою виконання завдання.

Завдання є важливою частиною роботи, яка повинна бути виконана в рамках бізнес-аналізу. Кожне завдання має бути виконано принаймні один раз протягом переважної більшості ініціатив бізнес-аналізу, але немає верхнього обмеження на кількість разів виконання будь-якого завдання[21].

Завдання можуть бути виконані в будь-якому масштабі. Кожне завдання може бути виконана протягом періоду від кількох місяців до кількох хвилин.

Наприклад, в якості бізнес-кейсу може бути документ з декількома сотнями сторінок, який обґрунтовує багатомільярдні доларові інвестиції, або одне речення, що пояснює вигоди зміни, яке буде вироблено для однієї людини[51].

Завдання має наступні характеристики:

- ✓ Завдання досягає результат в вихідних даних, який створює цінність для організації-спонсора, тобто якщо завдання виконується, то вона

повинна принести деякий очевидний позитивний результат, який корисний, конкретний, видимий та його можна оцінити.

- ✓ Завдання є завершеним, якщо завдання в перспективі можуть бути виконані різними людьми або групою, при використанні вихідних даних цього завдання.

Завдання є необхідною частиною призначення області знань, з якою вона пов'язана.

Упорядкування завдань є неминучим, тому що деякі завдання виробляють вихідні дані, які потрібні в якості вхідних даних для інших завдань. Тим не менш, важливо мати на увазі, що вхідні дані повинні існувати. Вхідні дані можуть бути неповними або можуть бути змінені або переглянуті, що може привести до виконання завдання кілька разів[33]. Ітераційний або гнучкий життєвий цикл може потребувати, щоб завдання у всіх областях знань виконувалися паралельно, а для життєвого циклу з чітко визначеними фазами буде як і раніше вимагатись, щоб завдання з багатьох областей знань були виконані на кожній фазі. Завдання можуть бути виконані в будь-якому порядку, за умови, що необхідні вхідні дані для завдання присутні.

Опис завдання більш детально пояснює, чому завдання виконується, що це за завдання і які результати завдання повинна досягти.

Вхідні дані надають інформацію та передумови необхідні для початку виконання завдання. Вхідні дані можуть бути такі що: явно генеруються за рамками бізнес-аналізу (наприклад, конструкція програмного додатка) або створюються за допомогою завдання бізнес-аналізу[49].

Немає припущення, що наявність вхідних і вихідних даних означає, що відповідний результат є в завершеному або в його фінальному стані. Ці дані повинні бути достатньо повними, щоб дозволити подальші роботи. Будь-яка кількість примірників вхідних даних можуть існувати протягом всього життєвого циклу ініціативи.

Вимоги є окремим випадком як вхідних або вихідних даних, що не повинно бути несподіванкою, враховуючи їх важливість для бізнес-аналізу. Вони є тільки

входом або виходом, які не створюються одним завданням. Вимоги можуть бути класифіковані в кілька різних способів і можуть існувати в будь-якому з безлічі станів.

Ефективність може бути визначена з точки зору зацікавлених осіб, які є одержувачами бізнес-аналізу. Усі зацікавлені особи можуть мати вихідні дані для оцінки цінності аналітичної роботи. Проте вхідні дані мають великі обсяги та зазвичай різнорідні та неструктуровані. Через це існуючі бізнес процеси не справляються в певній мірі зі своїми задачами.

У розрізі поставленої нами задачі: виявлення закономірностей і факторів, які впливають на лояльність абонентів, нам потрібно покращувати існуючи бізнес процеси.

2.4 Бізнес аналітика в телекомунікаціях

На тлі дедалі більшого поширення мобільних технологій операторам зв'язку необхідна актуальна, оперативна, досить докладна і персоніфікована інформація про абонентів, що дозволяє їх групувати, утримувати і ефективно обслуговувати. При цьому необхідна можливість виявляти патерни подій, як відбуваються в реальному часі, так і зареєстрованих протягом багатьох місяців - це дозволить завчасно попереджати виникнення проблем і виробляти заходи щодо підвищення якості обслуговування[47]. Отримання такої інформації вимагає обробки і аналізу великих обсягів різнорідних даних, що надходять з найрізноманітніших джерел - зі смартфонів, від датчиків, з мереж передачі даних і соціальних мереж, електронних листів, систем торгівлі цінними паперами та списків спостереження. При цьому виникає проблема реєстрації, фільтрації, очищення, організації, аналізу і подальшої обробки всіх цих інформаційних потоків, а також великих масивів історичних даних - з метою визначення прибутковості абонентів, ймовірності їх відтоку і контролю виконання умов обслуговування.

Один з найважливіших факторів ефективності бізнесу і збереження конкурентоспроможності - висока якість і актуальність інформації про абонентів. Для цього необхідно мати можливість виявляти зміни в поведінці

абонентів шляхом знаходження певних патернів в сукупності надходять з мережі даних про обслуговування, споживанні послуг і транзакціях[44].

Аналіз абонентської інформації виконується шляхом сегментації і побудови передбачуваних моделей, які дозволяють з'ясувати, які абоненти найбільш прибуткові, а які з найбільшою ймовірністю готові перейти до іншого оператору. Використання цих даних дозволяє визначити, які нові пакети послуг з найбільшою ймовірністю дозволять утримати абонентів, а також як вирішити проблеми якості обслуговування до того, як вони набудуть загрозливого масштабу.

Впровадження передових методів обробки абонентської інформації породжує для операторів зв'язку ряд проблем, пов'язаних з інфраструктурою обробки даних: збільшення числа користувачів, ускладнення запитів, затримки, пов'язані з обробкою великих масивів даних.

Інформаційно-обчислювальні потужності, необхідні для реалізації передових методів аналітичної обробки, значно перевершують граничні можливості експлуатованих в даний час программно-технічної інфраструктури - це так званий «розрив в аналітичних можливостях».

Великі обсяги даних. Операторам зв'язку доводиться обробляти набагато більше даних, ніж рік або навіть два роки тому. Це обумовлено появою смартфонів і мобільного широкосмугового доступу до мережі, обміну трафіком між клієнтами, а також зростанням споживання відео-сервісів. Додаткові фактори росту обсягів даних - необхідність поглибленого аналізу і підвищення точності передбачуваних моделей, при цьому потрібно реєструвати більше даних і зберігати їх за більш тривалий період.

Зростання кількості користувачів і пристроїв. Для підвищення якості обслуговування і зниження навантаження на абонентську службу ряд послуг зараз надається через Інтернет. Число користувачів, що звертаються до порталів обслуговування клієнтів, вельми високо. Оператори зв'язку повинні обслуговувати абонентів, які звертаються через Інтернет, з стійким якістю,

незалежно від обсягу трафіку, що приймається їх веб-додатками в кожен конкретний момент.

Складні запити. Більшість оперативних рішень і раніше приймається без комп'ютерної підтримки, що веде до суб'єктивізму, а також суперечить корпоративним правилам. В інших випадках логіка підтримки прийняття рішень жорстко «зашита» в системах BSS / OSS, що ускладнює її модифікацію з метою адаптації до мінливих потребам. Операторам необхідно за допомогою складних запитів порівнювати і зіставляти різні набори даних, виявляючи тенденції, причинно-наслідкові зв'язки і патерни.

Затримки, обумовлені великим обсягом даних. Ще один фактор, що ускладнює аналіз інформації про клієнтів - фактор часу. Для побудови якомога повнішої картини якості обслуговування оператори зв'язку повинні мати можливість отримувати практично цінну інформацію з розрізнених джерел мережі. Витяг інформації повинно виконуватися протягом декількох секунд після виникнення подій, а не через хвилини або години, як це відбувається сьогодні - таким чином, необхідний новий підхід.

Багато операторів усвідомили, що системи управління даними, впроваджені раніше для підтримки систем OSS / BSS, не задовольняють поточним вимогам до абонентської аналітиці реального часу. Застосування традиційних засобів бізнес-аналітики разом з реляційними базами даних часто призводить до незадовільних результатів, коли IT-служба змушена обмежувати число користувачів, мають доступ до даних, складність запускаються запитів або ж глибину ретроспективного аналізу записів про надані послуги[39].

Ці традиційні системи бізнес-аналітики та оперативні системи неефективні для реалізації передових методів аналітичної обробки реального часу по ряду причин. По-перше, єдиний можливий в цьому випадку режим роботи з транзакційними системами - просте отримання звітів. При цьому IT-служба змушена поміщати дані для аналізу і випуску звітів в безліч предметних сховищ. Для передбачення тенденцій шляхом ретроспективного аналізу даних про надані послуги необхідні додаткові IT-ресурси, що забезпечують витяг, перетворення і

завантаження даних. Зростання обсягів даних диктує необхідність в застосуванні зведених і агрегованих таблиць (для прискорення обробки запитів традиційними базами даних). Це ускладнює проекти зі створення сховищ даних і систем підтримки прийняття рішень і збільшує обсяг робіт.

Тому для того щоб покращити ефективність бізнес процесів в телекомунікаціях потрібно застосовувати машинне навчання. Machine learning в перспективі може дати змогу оброблювати великі данні з високою точністю та швидкістю. Розглянемо деталі машинного навчання та методи, які будуть використані для підвищення ефективності бізнес процесу.

2.5 Застосування методів машинного навчання

Машинне навчання (machine learning) - це метод аналізу даних, який автоматизує побудову аналітичної моделі. Це галузь штучного інтелекту, заснована на ідеї, що машини повинні вміти вчитися і адаптуватися через досвід. Воно тісно пов'язане з обчислювальної статистикою, яка робить прогнози на основі статистичних даних, зібраних комп'ютером [32].

Машинне навчання все глибше проникає в наше життя за допомогою призначених для користувача продуктів, створених за допомогою методів штучного інтелекту. очевидно, що дані технології будуть розвиватися і далі, поступово стаючи частиною повсякденної рутини в багатьох областях людської професійної діяльності. Однак з часів своєї появи, машинне навчання встигло обзавестися численними проблемами, головна з яких – досить висока трудомісткість. Побудова систем машинного навчання вимагає величезної кількості часу високопрофесійних фахівців як у сфері штучного інтелекту, так і в тій предметної області, до якої ця технологія застосовується.

Найбільш перспективною й актуальною в даний час технологією автори вважають автоматизоване машинне навчання - комплекс інструментальних і методичних засобів, що дозволяє значно скоротити частку людського участі в створенні систем штучного інтелекту, в тому числі засобами автоматичної валідації результатів моделювання.

Data Minig можна вважати надмножиною безлічі різних методів отримання даних із даних. Це може залучати традиційні статистичні методи та машинне навчання. Data Minig застосовує методи з багатьох різних областей для виявлення раніше невідомих закономірностей на основі даних. Це може включати статистичні алгоритми, машинне навчання, аналіз тексту, аналіз часових рядів та інші галузі аналітики. Data Minig також включає вивчення та практику зберігання даних та маніпулювання ними.

Основна відмінність від машинного навчання полягає в тому, що, як і статистичні моделі, мета полягає в тому, щоб зрозуміти структуру даних - пристосувати теоретичний розподіл до даних, які добре розуміються. Отже, із статистичними моделями існує теорія, яка стоїть за моделлю, яка математично доведена, але для цього потрібно, щоб дані також відповідали певним вагомим припущенням. Машинне навчання розроблено на основі здатності використовувати комп'ютери для зондування даних для структури, навіть якщо ми не маємо теорії того, як виглядає ця структура. Тест для моделі машинного навчання - це помилка перевірки нових даних, а не теоретичний тест, який доводить нульову гіпотезу. Оскільки машинне навчання часто використовує ітераційний підхід для вивчення даних, навчання може бути легко автоматизовано. Пропуски проходять через дані, поки не буде знайдено надійний шаблон.

Deep learning поєднує в собі досягнення обчислювальної потужності та спеціальні типи нейронних мереж для вивчення складних шаблонів у великих обсягах даних. Методи Deep learning на сьогодні є найсучаснішими для ідентифікації об'єктів на зображеннях та слів у звуках. Зараз дослідники прагнуть застосувати ці успіхи в розпізнаванні образів для більш складних завдань, таких як автоматичний переклад мови, медичні діагностики та численні інші важливі соціальні та ділові проблеми.

В телеком системах генерується надвелика кількість даних, які потрібно обробляти і аналізувати. Машинне навчання в основному використовує діапазон або спектр на основі методу оптимізації великої кількості параметрів.

Збільшення кількості та варіації доступних даних, не зважаючи на збільшення та різноманітність методів і засобів їх обробки, які стають більш дешевими і потужними, наявність більш доступних сховищ даних, заважає вирішенню проблеми аналітичної обробки в телеком індустрії. Таким чином, машинне навчання швидко стає дуже важливою і широко впроваджуваною частиною ділових процесів в системах телеком-оператора.

Запропоновано процес вирішення завдання методами машинного навчання. Він включає в себе певні етапи, проілюстровані на рис. 2.3 Процес машинного навчання може бути як успішним, так і неуспішним.

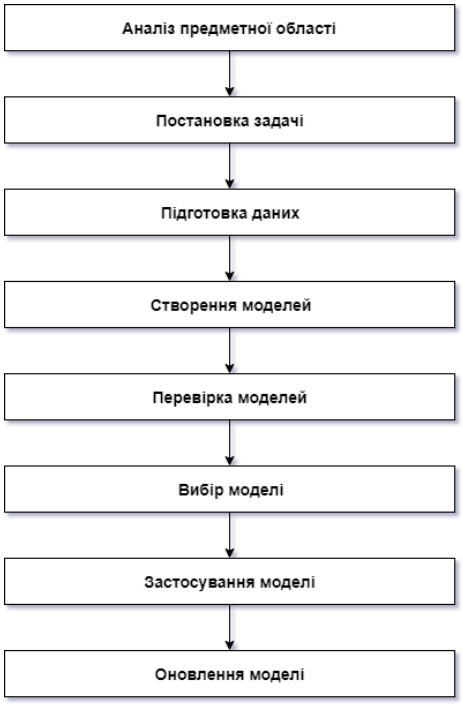

Рис. 2.3 Етапи виконання машинного навчання

Машинне навчання, як правило, поділяється на дві основні категорії: навчання з учителем та без учителя.

Метод з учителем в основному навчає машини на прикладах. Під час навчання для контрольованого навчання системи піддаються впливу великої кількості маркованих даних, наприклад, зображень рукописних фігур, анотованих, щоб вказати, якому номеру вони відповідають. Враховуючи достатньо прикладів, система контрольованого навчання навчиться розпізнавати скупчення пікселів та фігур, пов’язаних із кожним числом, і врешті-решт зможе розпізнавати рукописні числа, здатні надійно розрізняти числа 9 та 4 або 6 та 8.

Однак навчання цих систем, як правило, вимагає величезних обсягів маркованих даних, причому деякі системи повинні мати мільйони прикладів для засвоєння завдання.

На відміну від алгоритму з учителем, алгоритми без учителя із виявленням закономірностей у даних намагаються виявити подібність, яка розділяє ці дані на категорії. Прикладом може бути Airbnb, що об'єднує будинки, які можна взяти в оренду по сусідству, або Google News, які щодня об'єднують історії на подібні теми.

Алгоритми навчання без нагляду не призначені для виділення конкретних типів даних, вони просто шукають дані, які можна згрупувати за подібністю, або аномалії, які виділяються.

Розглянемо більш докладно методи машинного навчання за допомогою яких в роботі буде створено прогноз відтоку абонентів.

2.5.1 Алгоритм дерева рішень

Ухвалення рішення - це процес раціонального або ірраціонального вибору альтернатив, що має на меті досягнення усвідомлюваного результату. Один з методів автоматичного аналізу даних є дерева рішень. Перші ідеї створення дерев рішень відносяться до робіт Ховленда (Hoveland) і Ханта (Hunt) кінця 50-х років XX століття. Однак, основною роботою, що дала імпульс для розвитку цього напрямку, стала книга Ханта (Hunt, E.B.), Меріна (Marin J.) і Стоуна (Stone, P.J) «Experiments in Induction», що побачила світ у 1966 р [55].

Дерева рішень, що використовуються в Data Mining, бувають двох основних типів:

- ✓ Аналіз дерева класифікації, коли результат, що передбачається є класом, до якого належать дані;
- ✓ Регресійний аналіз дерева, коли результат, що передбачається, можна розглядати як дійсне число [5].

Data mining (видобування даних, інтелектуальний аналіз даних, глибинний аналіз даних) - збірна назва, що використовується для позначення сукупності методів виявлення в даних раніше невідомих, нетривіальних, практично

корисних і доступних інтерпретацій знань, необхідних для прийняття рішень в різних сферах людської діяльності. Термін введений Григорієм Пятецким-Шапіро у 1989 році. На сьогоднішній день існує велика кількість алгоритмів, що реалізують дерева рішень: CART, C4.5, CHAID, CN2, NewId, ITrule та інші.

Дерево прийняття рішень використовується в області статистики та аналізу даних для моделей, що прогнозуються. На ребрах дерева рішення записані атрибути, від яких залежить цільова функція, в вершинах записані значення цільової функції, а в інших вузлах - атрибути, за якими розрізняються випадки

Дерева рішень застосовуються в задачах класифікації (прийняття рішення про належність об'єкта до одного М з непересічних класів) і регресії (прогноз значення з безперервного діапазону). Класифікація і регресія на основі дерев рішень використовуються в задачах розпізнавання тексту, інформаційного пошуку, розпізнавання мови, аналізі зображень, виявленні спаму, розпізнавання жестів і ін. Для конструювання дерев рішень застосовується машинне навчання - автоматична настройка параметрів алгоритму на основі навчальної вибірки. При цьому від якості навчання залежить правильність рішення задачі і практична застосовність результатів [56].

Під алгоритмом будемо розуміти функцію, приймаючу на вхід класифікується об'єкт і повертає один з М класів - відповідь алгоритму для даного об'єкта. Дерева рішень складаються з вершин, в яких записані перевіряються умови (будемо називати ці умови ознаками), і листя, в яких записані відповіді дерева (один з М класів для завдання класифікації). Під навчальним прикладом будемо розуміти об'єкт навчальної вибірки з відомим правильною відповіддю (класом, до якого належить даний об'єкт). Навчання складається в налаштуванні умов у вузлах дерева і відповідей в його листі з метою досягнення максимальної якості класифікації [5].

Нехай задана кінцева множина об’єктів $X = \{x_1, \ldots, x_L\}$ і алгоритмів $A = \{a_1, \ldots, a_D\}$ і бінарна функція втрат $I: A \times X \to \{0, 1\}, I\{a, x\} = 1$ тоді і тільки тоді, коли алгоритм допускає помилку в обєкті *x*. Число помилок алгоритму *a* на вибірці *X* визначається як $n(a, X) = \sum_{x \in X} I\{a, x\}$. Частота помилок алгоритму

на вибірці визначається як $v(a,X) = n(a,x)/|X|$. Під якістю класифікації розуміється частота помилок алгоритму на контрольній вибірці.

Переваги та недоліки дерева рішень

Автоматичний відбір ознак. Ознаки в вершини дерева вибираються автоматично з набору ознак. Тому можна скласти довільний набір ознак, а в процесі навчання автоматично виберуться інформативні і проігнорують неінформативні ознаки. Немає необхідності в додатковій процедурі відбору ознак, на відміну від інших методів машинного навчання.

Інтерпритуємість. Дерева рішень дозволяють будувати вирішальні правила в формі, зрозумілою експерту. Це виявляється корисним в тому випадку, коли людині потрібно розуміти, яким чином алгоритм буде приймати рішення. Інтерпритуємість виявляється корисною властивістю, якщо потрібно зрозуміти, чому дерево рішень не працює належним чином.

Керованість. Якщо деякі приклади класифікуються неправильно, можна заново навчити тільки ті вершини дерева, через які це відбувається, що дуже зручно, коли обсяг навчальних даних великий і навчання займає багато часу. Крім того, при тренуванні різних під дерев можуть виявитися більш ефективними різні алгоритми навчання [55].

2.5.2 Алгоритм пошуку асоціативних правил

Алгоритми обмеженого перебору обчислюють частоти комбінацій простих логічних подій в підгрупах даних.

Асоціативні правила дозволяють знаходити закономірності між пов'язаними подіями. Такі правила формуються на підставі часто зустрічаються наборів даних [3].

Асоціативні правила прийнято представляти у вигляді наступного умовного судження [55]:

ЯКЩО (умова) ТО (результат),

де умова - набір об'єктів належать множині Z з якими асоційовані об'єкти, що входять в результат правила. Нехай умова = X, результат = У, тоді асоціативне правило можна представити в наступному вигляді:

якщо X то У ,

До головних переваг асоціативних правил відносять їх простоту сприйняття людиною і нескладні алгоритми.

Асоціативні правила можна розділити на наступні категорії:

- тривіальні правила - не несуть нової і практично корисної інформації, оскільки знання, які в них містяться, відомі і легко пояснити.
- незрозумілі правила - формуються на підставі яких глибоко прихованих знань, або аномальних даних (інформація, що міститься в них, не може бути пояснена логікою).
- нетривіальні (корисні) правила містять раніше невідомі і практично корисні знання (можуть бути пояснені логікою).

Щоб якісно оцінити отримані правила використовують такі величини:

- Підтримка (support) - показує, відсоток транзакцій, який підтримує дане правило. На підставі набору будується правило. Якщо декілька правил побудовані на підставі одного ж набору, то вони мають таку саму підтримку.
- Достовірність (confldence) - показує відсоток того, що з наявності в транзакції набору X слідує наявність в ній набору У. Чим достовірність більше, тим правило краще [32].

Алгоритм пошуку асоціативних правил призначений для знаходження всіх правил X Y, причому підтримка і достовірність цих правил повинні бути вищими за деякі наперед задані пороги, що називаються відповідно мінімальною підтримкою (minsupport) і мінімальною достовірністю (minconfidence). Завдання знаходження асоціативних правил розбивається на дві підзадачі [56]:

- Знаходження всіх наборів елементів, які задовольняють порогу minsupport. Такі набори елементів називаються такими, що часто зустрічаються.
- Генерація правил з наборів елементів, що знайдені згідно п.1. з достовірністю, яка задовольняє порогу minconfidence.

2.5.3 Метод «Bagging»

Bagging - метод класифікації, в якому всі елементарні класифікатори навчають і виконують роботу паралельно. Ідея є в тому, що класифікаторам не потрібно виправляти помилки один одного, а компенсація їх роботи відбувається при голосуванні [6]. Класифікатори що є базовими повинні бути незалежними один від одного, в якості цього можуть виступати класифікатори, що використовують групи методів або ж ті, що проходять навчання на наборах даних, що є незалежними. При навчанні на незалежних наборах, можна використати один метод.

Цей метод застосовується для класифікації багатовимірних об'єктів. Розглянутий метод допомагає домогтися якісної класифікації в умовах, коли розділити об'єкти на групи на всій множині параметрів не представляється можливим. Пропонується розділити простір характеристик на підмножини об'єднаних за змістом параметрів. Класифікація на кожній підмножині проводиться окремо, потім результати враховуються в голосуванні. У цьому випадку буде врахований внесок кожної смислової групи і підвищиться ймовірність того, що підсумкові результати класифікації виявляться більш якісними ніж без поділу на підмножини, так як параметри, за якими представники різних класів відрізняються, потраплять напевно, не в усі групи.

Існує матриця характеристик об'єкта $X: x_1, \dots, x_n$ - m-мірні стовпці з характеристиками n об'єктів. Необхідно зіставити кожному вектору параметрів мітку класу (тобто існує деяке відображення X-> Y, де $Y = (y_1 \dots y_k)$, y_i- мітки класів), на підставі відомих пар (x_i, y_j) для об'єктів навчальної вибірки.

Алгоритм класифікації в технології беггінг на підмножинах

1. Необхідно розділити простір параметрів на підмножини, тобто кожен об'єкт буде характеризуватися вже не одним m-мірним вектором параметрів, а кількома векторами $x_{i,1} \dots y_{i,l}$ причому сума розмірностей цих векторів не може перевищувати m, тобто підмножини не можуть перетинатися. Для цього вдаються до

експертної думки, експерт виділяє смислові підмножини на підставі свого досвіду.

2. Проводиться незалежне навчання кожного елементарного класифікатора (кожного алгоритму, визначеного на своїй підмножині).
3. Проводиться класифікація основної вибірки на кожній з підмножин (також незалежно).
4. Приймається остаточне рішення про належність об'єкта одному з класів. Це можна зробити декількома різними способами, докладніше описано нижче [6].

Остаточне рішення про належність об'єкта класу може прийматися, наприклад, одним з таких методів:

1. Консенсус: якщо всі елементарні класифікатори присвоїли об'єкту одну і ту ж мітку, то відносимо об'єкт до обраного класу.
2. Проста більшість: консенсус можна досягти дуже рідко, тому найчастіше використовують метод простої більшості. Тут об'єкту присвоюється мітка того класу, який визначило для нього більшість елементарних класифікаторів.
3. Зважування класифікаторів: якщо класифікаторів парна кількість, то голосів може вийти порівну, ще можливо, що для експерти одна з груп параметрів важлива більшою мірою, тоді вдаються до зважування класифікаторів. Тобто при голосуванні голос класифікатора множиться на його вагу.

Беггінг на підмножинах дозволяє згладжувати негативний вплив на загальну якість класифікації невіддільності класів за деякими параметрами. Метод потенційних функцій - це простий в реалізації метричний метод класифікації, який максимально ефективний тільки для компактних класів, проте, на відміну від, наприклад, методу січних площин, працює з класами різних форм (будь-яких).

2.6 Оцінка якості моделей і порівняння різних алгоритмів машинного навчання

У задачах машинного навчання для оцінки якості моделей і порівняння різних алгоритмів використовують такі метрики [55]:

- *Accuracy*,
- *precision*,
- *recall*,
- інтегрований показник *F-міра*.

Набір даних представляє собою таблицю розмірністю m, яка складається з параметрів p_i де $i=1,m$. Кожний i-тий параметр у рядку рі таблиці приймає певні значення. Таким чином кожен рядок таблиці відповідає k-ому, де k=1,n, стану процесу, який аналізують.

У найпростішому випадку такою метрикою може бути частка станів набору параметрів, за якими класифікатор прийняв правильне рішення [56].

$$Accuracy = \frac{P}{N} \qquad (2.1)$$

де, P – кількість станів набору параметрів, за якими класифікатор прийняв правильне рішення, N - розмір навчальної вибірки.

У цій метриці є одна особливість, яку необхідно враховувати. Вона полягає у призначенні всім параметрам однакової ваги, яка може бути некоректною у разі, якщо розподіл параметрів у навчальній вибірці є сильно зміщений у бік якогось одного або декількох класів. В цьому випадку у класифікатора є більше інформації стосовно цих класів і, відповідно, в рамках цих класів він буде приймати більш адекватні рішення. На практиці це призводить до того, що має місце неоднозначне визначення метрики accuracy для різних класів, розбіжність може сягати від 80% на певному класі до приблизно 0% на іншому[41].

Вихід з цієї ситуації полягає в тому, щоб навчати класифікатор на спеціально підготовленій, збалансованій вибірці групи параметрів. Недоліком цього рішення є втрата інформації щодо відносної частоти зміни значень параметрів.

ROC-крива (Receiver Operator Characteristic) - крива, яка найбільш часто використовується для представлення результатів бінарної класифікації в machine learning. Назва прийшла з систем обробки сигналів. Оскільки є два класи, один з них називається класом з позитивними наслідками, другий - з негативними наслідками. ROC-крива показує залежність кількості вірно класифікованих позитивних прикладів від кількості невірно класифікованих негативних прикладів. У термінології ROC-аналізу перша складова називається істинно позитивним, друга - хибно негативною множиною. При цьому передбачається, що у класифікатора є деякий параметр, варіюючи який, ми будемо отримувати те чи інше розбиття на два класи. Цей параметр часто називають порогом, або точкою відсікання. Залежно від нього будуть виходити різні величини помилок I і II роду [55].

У логістичної регресії поріг відсікання змінюється від 0 до 1 - це і є розрахункове значення рівняння регресії. Будемо називати його рейтингом.

Для розуміння суті помилок I і II роду розглянемо таблицю 2.1 спряженості (confusion matrix), яка будується на основі результатів класифікації моделі і фактичної (об'єктивної) належності прикладів до класів [56].

Таблиця 2.1

Таблиця спряженості

	Фактично	
Модель	Позитивно	Негативно
Позитивно	TP	FP
Негативно	FN	TN

TP (True Positives) - вірно класифіковані позитивні приклади (так звані істинно позитивні випадки);

TN (True Negatives) - вірно класифіковані негативні приклади (істинно негативні випадки);

FN (False Negatives) - позитивні приклади, класифіковані як негативні (помилка I роду). Це так званий "помилковий пропуск" - коли зацікавлені об’єкти помилково не виявляються (помилково негативні приклади);

FP (False Positives) - негативні приклади, класифіковані як позитивні (помилка II роду). Це помилкове виявлення, коли при відсутності події помилково виноситься рішення про його присутність (помилково позитивні випадки) [6].

При аналізі частіше оперують абсолютними показниками, а відносними - частками (rates), вираженими в відсотках. Частка істинно позитивних прикладів (True Positives Rate):

$$TPR = \frac{TP}{TP + FN} * 100\% \quad (2.1)$$

Частка хибно позитивних прикладів (False Positives Rate):

$$FPR = \frac{FP}{TN + FP} * 100\% \quad (2.3)$$

Введемо ще два визначення: чутливість і специфічність моделі. Ними визначається об'єктивна цінність будь-якого бінарного класифікатора.

Повнота (recall) - це і є частка істинно позитивних випадків:

$$recall = TPR = \frac{TP}{TP + FN} * 100\% \quad (2.4)$$

Точність (precision) - частка правильно прогнозованих екземплярів серед усіх знайдених:

$$precision = \frac{TP}{TP + FP} \quad (2.5)$$

Модель з високою чутливістю часто дає істинний результат при наявності позитивного результату (виявляє позитивні приклади) та навпаки, модель з високою специфічністю частіше дає істинний результат при наявності негативного результату (виявляє негативні приклади) [55] .

Зрозуміло що чим вище точність і повнота, тим краще. Але в реальному житті максимальна точність і повнота недосяжні одночасно і доводиться шукати якийсь баланс. Тому, хотілося б мати якусь метрику яка об'єднувала б у собі

інформацію про точність та повноту нашого алгоритму. У цьому випадку нам буде простіше приймати рішення про те, яку реалізацію запускати в production (у кого більше той і крутіше). Саме такою метрикою є F-міра [56].

F-міра є гармонійне середнє значення між точністю і повнотою. Вона наближається до нуля, якщо точність або наближається прагне до нуля.:

$$F = 2 * \frac{precision * recall}{precision + recall} \quad (2.6)$$

Дана формула надає однакову вагу точності і повноти, тому F-міра буде падати однаково при зменшенні і точності і повноти. Можливо розрахувати F-міру надавши різну вагу точності і повноти, віддаючи пріоритет однієї з цих метрик при розробці алгоритму.

$$F_\beta = (1 + \beta^2) * \frac{precision * recall}{(\beta^2 * precision) + recall} \quad (2.7),$$

де β приймає значення в діапазоні 0 <β <1 якщо ви хочете віддати пріоритет точності, а при β> 1 пріоритет віддається повноті. При β = 1 формула зводиться до попередньої і ви отримуєте збалансовану F-міру (також її називають F1).

ROC крива - Receiver operating characteristic, це графік, що дозволяє оцінити якість поділу двох класів. Крім візуальної складової, є чисельна характеристика ROC AUC - Area under ROC curve (AUROC, ROC AUC), площа під ROC кривої, чим вище значення - тим краще, 0.5 - погана класифікація, яка не відрізняється від рівно вірогідних, більше 0.75 - вважається хороша класифікація, і більше 0.8 - вже найкраща класифікація, якщо значення 0.2 - то необхідно повторити експеримент заново і знайти допущенні помилки. ROC curve є графіком двох значень: співвідношення кількості правильно і неправильно класифікованих ознак [55].

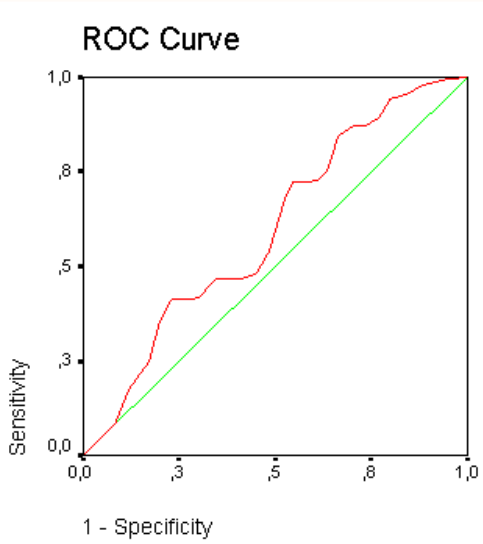

Рис 2.4 Приклад ROC-кривої

Завдання класифікації полягає в тому, щоб відносити раніше невідомі суті до того чи іншого клас. В цьому випадку є два класи результатів: позитивний (positive) і негативний (negative). Тоді на виході класифікатора може спостерігатися чотири різних ситуації [56]:

- Якщо результат класифікації позитивний, і справжнє значення теж позитивний, то мова йде про істинно-позитивному значенні (true-positive, TP).
- Якщо результат класифікації позитивний, але справжнє значення негативне, то мова йде про хибно-позитивному значенні (false-positive, FP).
- Якщо результат класифікації негативний, і справжнє значення теж негативне, то мова йде про істинно-якого від'ємного значення (true-negative, TN).
- Якщо результат класифікації негативний, але справжнє значення позитивно, то мова йде про хибно-негативному значенні (false-negative, FN)

Висновки

В даному розділі розглянуто методи проведення бізнес аналізу в телекомунікаційних системах. Результати бізнес процесу аналітичної обробки інформації не завжди відповідають очікуванням через великих обсяг

неструктурованих даних, які генеруються на вході системи, тому оператор не завжди отримує достовірний результат аналізу. Тому існуючи методи бізнес аналізу потребують удосконалення.

Для удосконалення бізнес процесів аналітичної системи телеком оператора пропонується застосовувати методи машинного навчання.

Аналіз математичних методів машинного навчання дозволив зробити висновки, що для вирішення задачі виявлення абонентів, які схильні до відтоку, пропонується створити комплексний метод до складу якого входять: метод дерева рішень, метод асоціативних правил та bagging.

РОЗДІЛ 3

СТВОРЕННЯ СЦЕНАРІЮ БІЗНЕС ПРОЦЕСУ ДЛЯ ЗАПОБІГАННЯ ВІДТОКУ АБОНЕНТІВ НА ОСНОВІ АНАЛІЗУ МЕТОДІВ МАШИННОГО НАВЧАННЯ

3.1 Етапи машинного навчання.

Розглянемо етапи збору та аналізу даних за допомогою методів машинного навчання. Важливо пройти всі етапи підготовки, обробки та аналізу даних для досягнення поставленої задачі.

Запропонована модель процесу машинного навчання в рамках поставленої нами задачі. Наглядно всі етапи представлені на Рис.3.1.

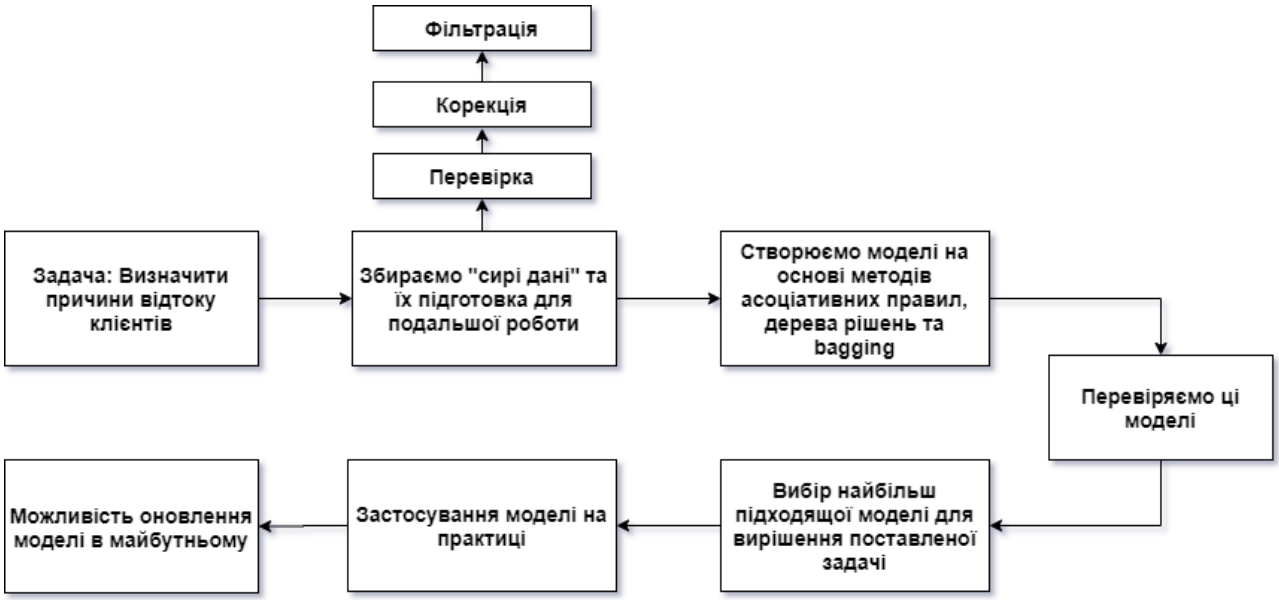

Рис. 3.1 Етапи виконання машинного навчання в рамках поставленої нами задачі

На початку роботи потрібно провести аналіз телекомунікаційної структури. Дані, які потім будуть оброблятися, збираються зі мережевого обладнання. Далі ставимо основну задачу, яку хочемо вирішити а саме: Визначити причини відтоку клієнтів, які саме фактори впливають на це, які складові даних мають найбільшу значимість. Якщо це все визначити, то в подальшому можна вплинути на визначені фактори та запобігти відтоку клієнтів.

Після того як визначено головну задачу, та зібрано дані, їх потрібно структурувати. Для цього вхідні дані повинні пройти наступні етапи: перевірка, корекція, фільтрація, розділення, маршрутизація.

Після того як це було зроблено, то переходимо до етапу моделювання.

3.2 Етап підготовки даних

Вхідні дані, які отримані під час роботи однієї української телекомунікаційної компанії, яка забезпечує абонентам мобільний зв’язок. Спочатку потрібно структурувати отримані дані. В таблиці даних присутні пропуски, що є недопустимим для аналізу та можуть ускладнити процес навчання системи. Пропуски можуть вносити додаткові похибки під час передбачення.

ABON_CODE	STATUS	COUNT_DAYS_OVER_1MB	COUNT_DAYS_OVER_5MB	DUAL_SIM_PROBABILITY	SIM_PRIORITY	OBLAST	CITY
1	0			High	SECOND	Donets'ka	Rodyns'ke
2	0			Very high	SECOND	Kharkivs'ka	Izium
3	0			Very high	SECOND	L'vivs'ka	Novoiavorivs'k
4	0			High	SECOND	Kharkivs'ka	Kharkiv
5	0	10	9	Very high	SECOND	Kharkivs'ka	Bohodukhiv
6	0	23	19	Very high	FIRST	Donets'ka	Makiivka
7	0	27	23	Low	FIRST	Kyivs'ka	Kyiv
8	0			Very high	SECOND	Kharkivs'ka	Kharkiv
9	0			Very high	FIRST	Kharkivs'ka	Bilyi Kolodiaz'
10	0			High	SECOND	Kharkivs'ka	Zavhorodnie
11	0			Very high	SECOND	Donets'ka	Artemivs'k
12	0			Very high	SECOND	L'vivs'ka	Kolodentsi
13	0			UNDEF	UNDEF	L'vivs'ka	Khodoriv
14	0			Low	FIRST	Odes'ka	Shyroka Balka
15	0			UNDEF	UNDEF	Kyivs'ka	Kyiv
16	0			Very low	FIRST	L'vivs'ka	Boryslav
17	0			Low	FIRST	Ternopil's'ka	Ternopil'
18	0			High	SECOND	Donets'ka	Makiivka
19	0			High	SECOND	Kyivs'ka	Kyiv
20	0			High	FIRST	Kharkivs'ka	Artil'ne
21	0	6	4	Very high	SECOND	Ternopil's'ka	Romanove Selo
22	0	2		Very high	SECOND	Zhytomyrs'ka	Berdychiv
23	0	4	3	Very low	FIRST	Zakarpats'ka	Chop
24	0			Very high	SECOND	L'vivs'ka	Hirne
25	0			Very high	SECOND	Rivnens'ka	Sarny
26	0	8	7	Low	FIRST	Rivnens'ka	Zdolbuniv
27	0			Very low	FIRST	L'vivs'ka	L'viv
28	0			Very high	SECOND	Mykolaivs'ka	Mykolaiv

Рис. 3.2 Вхідні дані

Для коректного виконання процесу аналізу набору даних потрібно заповнити пропуски, які присутні у вхідних даних (рис. 3.2.). Для цього використовуємо функцію пошуку пропусків та заповнюємо їх нулями. Таким чином під час моделювання отримуємо більшу вірогідність правдивих результатів.

	STATUS	COUNT_DAYS_OVER_1MB	...	INET_SLOPE	REFILL_SLOPE
0	0	0.0	...	0.000000	0.000000e+00
1	0	0.0	...	0.022223	-4.350000e-36
2	0	0.0	...	0.046196	3.710257e-02
3	0	0.0	...	-0.026511	3.838383e-02

Рис. 3.3 Підготовлені вхідні дані для передбачення

3.3 Моделювання та передбачення відтоку абонентів

При проведенні моделювання поведінки клієнта з метою передбачення відтоку абонентів використано підхід, представлений в роботі [52]. В статті застосовано метод випадкового лісу, за допомогою якого проведено моделювання. Випадковий ліс - один із прикладів об'єднання класифікаторів в ансамбль. Моделювання виконувалось на мові програмування Python. В статті визначено вибірку абонентів, які схильні до відтоку з заданими ймовірностями.

Таблиця 3.1

Абоненти, які схильні до відтоку

№	prob_true
1977	0.9333
2696	0.9167
2708	0.9233
2924	0.9041

В роботі застосовано методи машинного навчання: асоціативних правил, дерева рішень та bagging з метою підвищення ефективності передбачення, збільшивши точність ймовірностей схильності до відтоку абонентів.

На першому етапі навчаємо систему за допомогою вищезгаданих методів машинного навчання на існуючих даних з уже відомим результатом відтоку (метод машинного навчання з учителем). Після того як система навчилась, на вхід системи надходять дані на обробку без відомого результату. Результат, який дала система звіряємо з уже відомим даними про відтік абонентів і рахуємо точність. Результати проведеного моделювання представлені далі.

Проводячи подальше моделювання над даними та використовуючи метод асоціативних правил отримуємо наступні результати:

Accuracy: 0.7319

Таблиця 3.2

Розбиття множини у відповідності до асоціативних правил

	Predicted False	Predicted True
Actual False	1791	737
Actual True	677	1774

Під час моделювання метод асоціативних правил показав наступні результати, які відображено в табл. 3.2. В даній таблиці показано кількість вірно та невірно оцінених даних. Показано помилки першого і другого порядку для методу асоціативних правил (методу машинного навчання). За допомогою цих даних було побудовано графік залежності ROC, який відображено в рис 3.4.

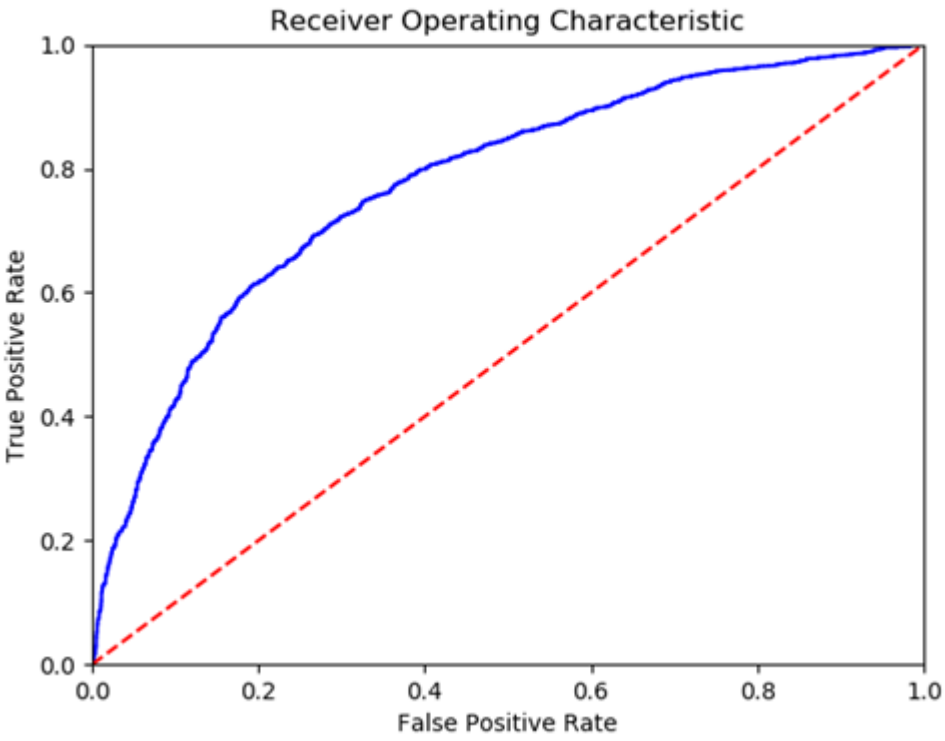

Рис. 3.4 Графік ROC-кривої для методу асоціативних правил

Далі було проведено моделювання, використовуючи метод дерева рішень і отримано наступні результати:

Accuracy: 0.7745

Таблиця 3.3

Розбиття множини у відповідності з методом дерева рішень

	Predicted False	Predicted True
Actual False	2065	475
Actual True	634	1827

Під час моделювання метод дерева рішень показав наступні результати, які відображено в табл. 3.3. В даній таблиці показано кількість вірно та невірно оцінених даних. Показано помилки першого і другого порядку для методу Data Mining, а саме для методу дерева рішень. За допомогою цих даних побудуємо графік залежності ROC, який відображено на рис 3.5.

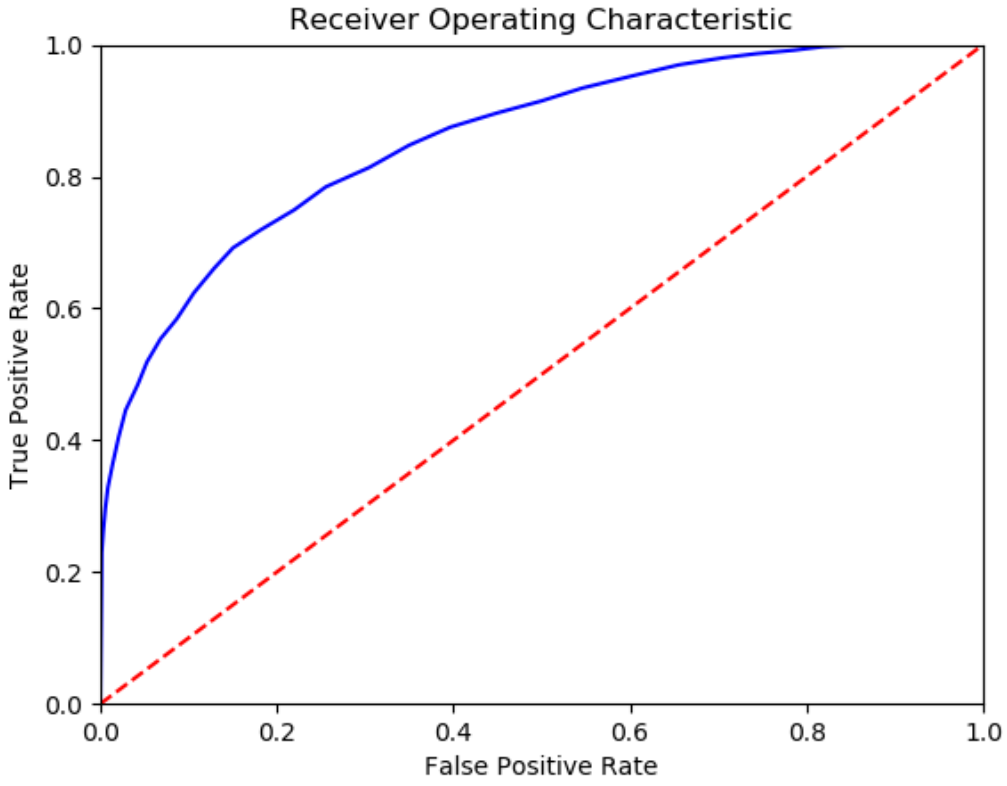

Рис. 3.5 Графік ROC-кривої для методу дерева рішень

Потім проведено моделювання, використовуючи метод Bagging і отримано наступні результати:

Accuracy: 0.8045

Таблиця 3.4

Розбиття множини у відповідності з методом bagging

	Predicted False	Predicted True
Actual False	2109	493
Actual True	664	1732

Під час моделювання метод bagging показав наступні результати, які відображено в табл. 3.4. В даній таблиці показано кількість вірно та невірно оцінених даних. Показано помилки першого і другого порядку для методу Data Mining, а саме для методу дерева рішень. За допомогою цих даних побудуємо графік залежності ROC, який відображено на рис 3.6.

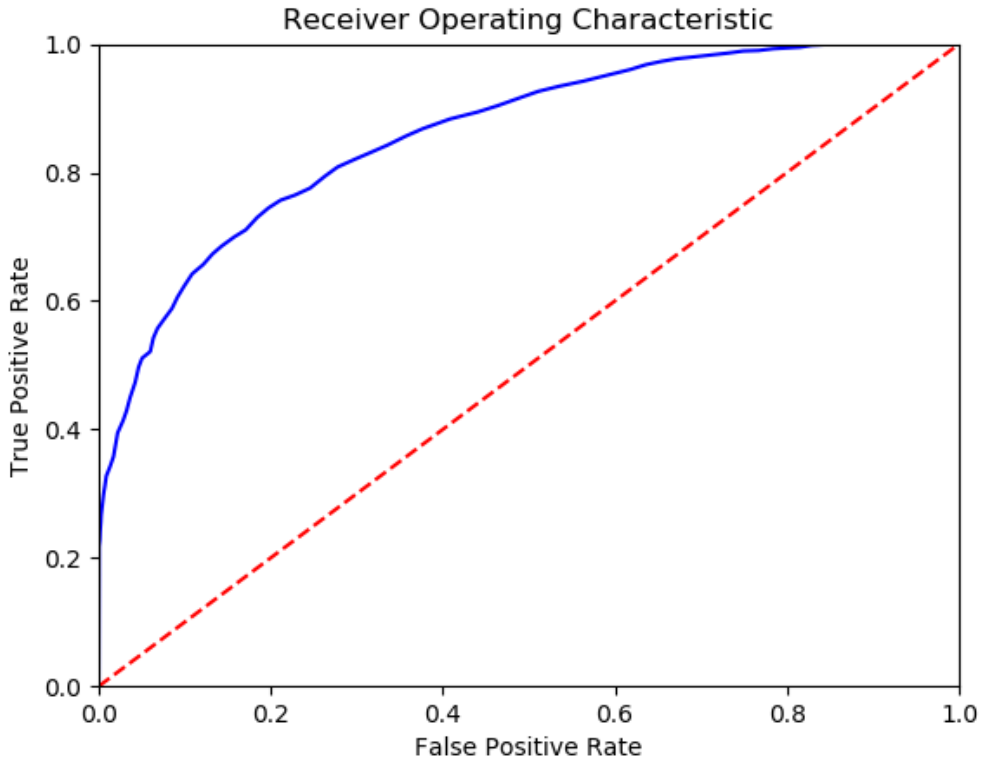

Рис. 3.6 Графік ROC для Bagging

Наступним етапом є виявлення факторів, що найбільше впливають на незадоволення абонента користування послугами телекомунікаційної компанії. Під час моделювання виявлено наступні фактори, які найбільш впливають на відтік клієнтів, їх представлено в табл. 3.5. Наступними факторами є:

- тривалість користування послугами телекомунікаційної компанії;
- кількість днів, під час якого абонент не користується послугами;
- загальна кількість дзвінків інших мобільних операторів;
- вихідні дзвінки на номери інших мобільних операторів;
- загальна кількість міжнародних дзвінків.

Таблиця 3.5

Вплив параметрів на процес відтоку клієнтів

№	Важливість	Назва
32	0.077898	Тривалість користування послугами телекомунікаційної компанії
23	0.071456	Кількість днів, під час якого абонент не користується послугами
21	0.061345	Загальна кількість дзвінків інших мобільних операторів
22	0.054581	Вихідні дзвінки на номери інших мобільних операторів
19	0.046387	Загальна кількість міжнародних дзвінків

Під час моделювання було виявлено абонентів, які найбільш схильні до переходу на послуг іншої телекомунікаційної компанії. В табл.3.6. наведена вибірка абонентів, які схильні до відтоку від оператора мобільного зв’язку. Поруч з кожним абонентом показано відсоток ймовірності відтоку. Можна зробити висновок, що на основі розгляду цих даних є можливість вплинути на окремих користувачів та надати кожному з них ті послуги, в яких абонент зацікавлений більше всього. Таким чином можна запобігти відтоку клієнтів.

Таблиця 3.6

Абоненти, які схильні до відтоку

№	prob_true
14692	0.99256
12897	0.99121
19630	0.98286
13752	0.95789
9579	0.94865

На основі значень ймовірності відтоку абонентів телекомунікаційної мережі була запропонована класифікація, яка представлена в таблиці 3.7.

Відповідно до класифікації всі абоненти поділені на групи, що мають відповідну ймовірність відтоку. Завдяки даній інформації оператори зв’язку мають змогу запобігти відтік клієнтів ввівши до свого бізнес плану правки, що беруть до уваги групи абонентів, що мають найбільші показники схильності до відтоку («високий» та «дуже високий» показники ймовірності відтоку абонента) та запропонувавши найбільш вигідні умови співпраці для даних груп, що в свою чергу задовільнить потреби абонента.

Таблиця 3.7

Класифікація абонентів відповідно до схильності відтоку

Показник ймовірності відтоку абонента	Відсоток
Дуже малий	До 20%
Малий	До 50%
Середній	До 70%
Високий	До 90%
Дуже високий	Більше 90%

Результати моделювання та передбачення щодо вирішення поставленої задачі, а також аналіз цих параметрів на точність та достовірність для різних

алгоритмів представлені в таблиці 3.8. Аналіз результатів дозволяє зробити наступні висновки: серед обраних методів найкращий результат показує метод Bagging, результати якого до 7% кращі ніж значення метрики методу асоціативних правил та до 3% кращі ніж значення метрики методу дерева рішень. Він перевершує метод асоціативних правил своєю повнотою, точністю та правдивої вірогідності прогнозування та перевершує метод дерева рішень швидкістю оброблення даних.

Таблиця 3.8

Значення метрик, які отримані для розглянутих методів

Модель	precision	recall	F1	F0.5	Accuracy
Асоціативних правил	0,710	0,733	0,718	0,711	0.7319
Дерева рішень	0,803	0,75	0,776	0,792	0.7745
Bagging	0,825	0,759	0,776	0,799	0.8045

На основі проведеного моделювання складемо порівняльну таблицю та побачити переваги та недоліки розглянутих методів. Результати порівняння представлені в табл. 3.9

Таблиця 3.9

Порівняння розглянутих методів машинного навчання

Метод	Переваги	Недоліки
Асоціативних правил	• Простота • Нескладні алгоритми • Швидко оброблюють невеликі обсяги даних	• При великих об’ємах даних дає досить велику ймовірність помилки
Дерево рішень	• Має здатність ефективно обробляти дані з великим числом ознак і класів	• Потрібно більше часу для обробки даних
Bagging	• Метод допомагає досягти якісної класифікації абонентів за рахунок поділу множини на підмножини	• У разі надвеликих надлишкових вибірок доводиться будувати підвибірки меншої довжини

За допомогою виявлених переваг і недоліків методів машинного навчання можемо навчити систему обирати доцільний метод в певній ситуації (залежність від вхідних даних).

1.4 Побудова сценарію бізнес процесу на основі розглянутих методів машинного навчання

Існує багато методів машинного навчання, які можуть вирішити задачу передбачення лояльності абонента. В різних випадках може виникнути ситуація в якій буде доцільно використовувати саме один конкретний метод серед інших. Такі ситуації можуть бути обумовлені різнорідністю та неструктурованістю інформації, яка подається на вхід аналізу передбачення відтоку абонентів.

На рис 3.7 запропонована модель в якій обирається один з запропонованих методів в залежності від даних, які поступають на вхід системи. Таким чином можна досягти ефективного використання системи застосовуючи доцільні методи в певних випадках.

Рис. 3.7 Запропонований сценарій бізнес процесу для вирішення задачі визначення абонентів, які схильні до відтоку

Спочатку беремо тестові дані, в яких вже відомо результат відтоку. В цих даних уже історично сформовані результати. Далі система навчається за допомогою обраних методів та визначаємо закономірності, від яких залежить

точність отриманих результатів передбачення (проводимо машинне навчання з учителем).

Після навчання система отримує вхідні дані з невідомим результатом лояльності абонентів. В залежності від об’ємів і якості вхідних даних (кількість абонентів та кількість факторів, які впливають на відтік абонентів) кожен з методів покаже різний результат. Тому потрібно проаналізувати дані, які отримані для передбачення відтоку та в залежності від кількості абонентів і факторів прийняти рішення щодо використання одного методу машинного навчання, який покаже найвищу точність за заданих умов. Для цього детально розглянемо блок аналізу даних рис. 3.8 .

Рис. 3.8 Аналіз вхідних даних

Неоднорідність та неструктурованість великих даних значно погіршують результат прогнозування, тому нам потрібно провести аналіз вхідних даних. Для того, щоб визначити, який саме метод нам потрібно застосувати, дані повинні пройти наступні етапи аналізу:

- Перевірка - дозволяє обмежити тип даних або значень, які знаходяться у вхідних даних.

- Корекція - метод надлишкового шумостійкого кодування / декодування каналу, що дозволяє виправити помилки введення. Застосовується для виправлення помилок та помилок шляхом введення надлишкової службової інформації до даних, яка може бути використана для відновлення вихідного вмісту вхідних даних..
- Консолідація - діяльність, спрямована на оптимальну організацію бази даних, яка реалізує всі необхідні взаємозв'язки між елементами даних, але база даних не містить дублікатів та зайвих елементів.
- Фільтрація - це процес пошуку і вибору даних відповідно до встановлених критеріїв. Фільтри також спрощують процес введення та видалення інформації із даних. При фільтрації дані, які не відповідають зазначеним критеріям, приховуються, але їх порядок розміщення в таблиці залишається незмінним і вони не вилучаються з загальних даних.
- Розділення - механізм, що дозволяє розділяти інформаційну базу на кілька областей, кожна з яких доступна тільки певним групам користувачів. При цьому зберігання і обробка даних у всіх областях виконується однією і тією ж конфігурацією.
- Представлення - це частина конфігурації вхідних даних в обліковому запис, що має окремі налаштування. Для даних можна створити кілька представлень інформації і налаштувати в кожному з них показ різної інформації про вхідні дані.
- Розподіл - сукупність логічно взаємопов’язаних баз даних, розподілених у комп’ютерній мережі. Логічне з'єднання баз даних у розподіленій базі даних забезпечується системою управління розподіленою базою даних, яка дозволяє управляти розподіленою базою даних таким чином, щоб створити ілюзію єдиної бази даних для користувачів..

Після того як пройдено етап аналізу, в результаті отримуємо структуровані дані та система може вирішити який метод доцільніше використати. Наприклад,

якщо в результаті на вхід системи надійде невеликий обсяг даних, тоді доцільніше всього буде застосувати метод асоціативних правил, адже цей метод дає достовірний результат за короткий проміжок часу (на відміну ніж інші методи).

Для того, щоб обрати доцільний метод, система звертає увагу на характеристики данних, які отримано на вході та на вимоги до прогнозування, а саме:

- Об’єм вхідних даних(терабайти, єксабайти)
- Точність отриманого прогнозування (достовірність прогнозування)
- Швидкість прогнозування (наскільки швидко дані будуть обробляютися)

На основі вищевказаних характеристик та методів машинного навчання, які було розглянуто у роботі, складемо матрицю, за якою система буде обирати дані в залежності від вимог на поточний момент (об’єм вхідних даних та точність/щвидкість прогнозування) табл.3.10.

Таблиця 3.10

Матриця прийняття рішень

Параметр	Об’єм вхідних даних	Точність прогнозування	Швидкість прогнозування
Об’єм вхідних даних	Дерева рішень	Bagging	Дерева рішень
Точність прогнозування	Bagging	Bagging	Асоціативних правил
Швидкість прогнозування	Дерева рішень	Асоціативних правил	Асоціативних правил

Матриця показує який саме метод потрібно обрати в тому чи іншому випадку. Наприклад, якщо на вхід системи надійшов великий об’єм даних та потрібно отримати найбільш точний результат - обираємо метод Bagging (базуючишь по створеній матриці). В подальшому ця матриця може удосконалюватись за рахунок додавання характеристик, яким повинні

відповідати вхідні данні або отримане прогнозування та додавання нових методів машинного начання. Таким чином система буде ставати більш гнучкою та зможе ефективно адаптуватися під відповідні умови.

В результаті отримано прогноз лояльності абонентів ефективним способом, використавши доцільний метод. Таким чином застосувавши різні методи машинного навчання отримано ефективний сценарій бізнес процесу для вирішення задачі передбачення відтоку клієнтів.

Висновки

В роботі проведено аналіз баз даних персоніфікованого трафіку алгоритмами пошуку асоціативних правил, дерев рішень та bagging з метою визначення лояльності потенційного клієнта, який показав, що зазначені методи дозволяють отримати практично ідентичні результати.

Прогнози алгоритмів відрізняються високою достовірністю, мають великі прогностичні можливості.

Поведено математичне моделювання методів передбачення відтоку клієнтів за допомогою методів асоціативних правил, дерева рішень та bagging та виявлено, що метод bagging дозволяє отримати більш точні результати, на 7% в порівнянні з методом асоціативних правил та до 3% кращі ніж значення метрики методу дерева рішень, точність отриманих результатів складає 80,45%.

Визначено фактори, які найбільш впливають на рішення абонента відмовитись від послуг оператора зв’язку, а саме:

- тривалість користування послугами телекомунікаційної компанії;
- кількість днів, під час якого абонент не користується послугами;
- загальна кількість дзвінків інших мобільних операторів;
- вихідні дзвінки на номери інших мобільних операторів;
- загальна кількість міжнародних дзвінків.

В певних умовах один з методів може виявитися більш ефективним, ніж інші. Для найбільш ефективного моделювання передбачення відтоку клієнтів

варто застосовувати ансамблеві рішення, що об’єднують розглянуті методи машинного навчання.

Було запропоновано комплексний метод, який вирішує задачу відтоку абонентів, використовує декілька методів машинного навчання, що дозволяє застосувати як кінцеве рішення найбільш ефективне за критеріями: точність, швидкість обробки та обсяг даних, що обробляються, дозволяючи підвищувати ефективність виявлення закономірностей і факторів, що впливають на лояльність абонента до телекомунікаційної компанії.

РОЗДІЛ 4
РОЗРОБЛЕННЯ СТАРТАП ПРОЕКТУ

В даному розділі викладено маркетинговий аналіз перспектив реалізації системи передбачення лояльності абонентів, а також оцінено можливості її впровадження. Розділ створено на основі методичних рекомендацій [57].

4.1 Опис ідеї проекту

Проект направлений на підвищення виявлення факторів та закономірностей, що впливають на рішення абонентів припинити користуватися послугами телеком оператору. Розроблена система допомагає операторам мобільного зв'язку утримувати клієнтів та впливати на їх лояльність, збирати та аналізувати інформацію про можливий відтік та визначати фактори, які мають найбільший вплив на відтік.

Таблиця 4.1

Опис ідеї стартап-проекту

Зміст ідеї	Напрямки застосування	Вигоди для користувача
Система визначення лояльності абонентів, які користуються послугами мобільного зв'язку, що дає можливість вплинути на відтік абонентів.	• Визначення лояльності абонентів. • Передбачення відтоку абонентів. • Визначення параметрів та факторів, які найбільше впливають на рішення абонента.	Оператори мобільного зв'язку можуть визначити вибірку абонентів, які найбільш схильні до відтоку і вплинути на їх рішення застосовуючи программи підвищення лояльності абонентів.

Під час порівнянні з конкурентними системами, в першу чергу звертають увагу на архітектуру системи, що забезпечує швидкість обробки даних та ефективність обраного методу машинного навчання у кожному конкретному

випадку. Порівняння з конкурентами, а також визначення переваг і недоліків наведено у наступні таблиці.

Таблиця 4.2

Визначення сильних, слабких та нетральних характеристик ідеї проекту

No п/п	Техніко-економічні характеристики ідеї	(потенційні) товари/концепції конкурентів			W (слабка сторона)	N (нейтральна сторона)	S (сильна сторона)
		Мій проект	SAS	Oracle Big data			
1	Використання безкоштовного програмного забезпечення	так	ні	ні			так
2	Непередбачуване зростання вартості	так	ні	ні	так		
3	Індивідуальні можливості кастомізації системи	так	так	ні			так
4	Єдина підтримка апаратної та програмної частини	ні	ні	так			так

Аналіз конкуретно спроможності дозволяє зробити висновок, що конкуренти мають лише частковий функціонал, який реалізований в нашому проекті. Серед сильних сторін визначені використання безкоштовного програмного комплексу та можливості кастомізації під конкретні задачі, які

ставляться перед обраним проектом. Слабкою стороною є вимога використовувати дані отримані з різних джерел, що не є структуровані.

4.2 Технологічний аудит ідеї проекту

Таблиця 4.3

Технологічна здійсненність ідеї проекту

No п/п	Ідея проекту	Технології реалізації	Наявність технологій	Доступність технологій
1	Система визначення лояльності абонента	Мова програмування Python	Існуючі бібліотеки, такі як Pandas, Numpy, Sklearn та Matplotlib що дозволяють визначити параметри відтоку	Є доступними та безкоштовними для використання
2		Мова програмування R	Фреймворки та бібліотеки, що дозволяють визначити параметри відтоку	Є доступними та безкоштовними для використання
Обрана технологія реалізації проекту: мова програмування Python				

Для реалізації проекту доступні дві мови програмування: Python та R.

Python використовується для вирішення великого обсягу задач, в різних випадках за рахунок великої кількості бібліотек та легкості використання. R являє собою більш спеціалізовану мову програмування для проведення аналізу великих даних (проведення аналітики великих об’ємів даних). Мову програмування для реалізації системи виявлення лояльності клієнтів було обрано Python.Завдяки можливості детальної документації та одночасне

використання коду великою кількостю користувачів, що забезпечують підтримку відкритих проектів, ця платформа є найбільш доцільною у використанні для реалізації проекту.

4.3Аналіз ринкових можливостей запуску стартап-проекту

Під час дослідження ринкових можливостей проекту, проведемо аналіз попиту: наявність попиту, обсяг, динаміка розвитку ринку. Дані представлені у таблиці нижче.

Таблиця 4.4

Попередня характеристика потенційного ринку стартап-проекту

No п/п	Показники стану ринку (найменування)	Характеристика
1	Кількість головних гравців, од	5
2	Загальний обсяг продаж	?
3	Динаміка ринку (якісна оцінка)	Зростає
4	Наявність обмежень для входу	Немає
5	Специфічні вимоги для стандартизації, специфікації	Немає
6	Середня норма рентабельності в галузі, %	?

Враховуючи сьогоднішню необхідність ринку рішень стосовно визначення лояльності клієта, за попереднім оцінюванням ринок є приваблим для впровадження.

Таблиця 4.5

Характеристика потенційних клієнтів стартап-проекту

No п/п	Потреба, що формує ринок	Цільова аудиторія (цільові сегменти ринку)	Відмінності у поведінці різних потенційних цільових груп клієнтів	Вимоги споживачів до товару
1	Утримання клієнтів	Оператори зв’язку	Оператори зв’зку в свою чергу можуть визначати самостійно формат взаємодії та внести свої коригування.	• Впровадження систем утримання клієнтів для зменшення відтоку. • Можливість отримати перелік параметрів котрі оказують найбільший вплив на відтік клієнтів. • Можливість отримати вибірку абонентів, які мають намір відмовитися від послуг мобільного оператора.

Таблиця 4.6

Фактори загроз

No п/п	Фактор	Зміст загрози	Можлива реакція компанії
1	Відсутність підтримки продукту	Успіх впровадження системи залежить від підтримки малого бізнесу, адже великі всеукраїнські оператори зв'язку не будуть звертати увагу на нового гравця на ринку аналізу великих даних та рішення проблеми передбачення відтоку абонентів, а виберуть вже перевірені часом рішення.	Робота з малим бізнесом над підвищенням визначення лояльності абонента, що буде означати створення бази клієнтів, які будуть рекомендувати цей продукт.
2	Великий об'єм	Найбільшою проблемою при роботі з великими даними є їх різнорідність та неструктурованість, а саме: пропуски, надлишковість, різна структура даних. Це все призводить до того що ще до того як почати аналізувати та робити висновки, потрібно провести великий обсяг роботи для структуризації цих даних та отримання відповіді на питання: «Мають ці данні корисну інформацію для нас чи ні?».	Використання методів математичного аналізу даних, використання методів машинног навчання та комбінація різних алгоритмів для отримання найкращого результату.

Таблиця 4.7

Фактори можливостей

No п/п	Фактор	Зміст можливості	Можлива реакція компанії
1	Необхідність рішення для проблеми передбачення та запобыгання відтоку клієнтів.	На сьогоднішній день телекомунікаційні компанії зазнають великих збитків через відтік абонентів. Для залучення нових клієнтів необхідно використовувати набагато більше ресурсів ніж для утримання клієнта	Просування продукту на всеукраїнський ринок з орієнтуванням на операторів мобільного зв’язку.
2	Зростання об’ємів даних про абонентів що користуються послугами телеком компанії	При роботі мобільного оператору виникає необхідність збору великих даних, котрі неможливо обробити старими методами або в ручну.	Висвітлення найпривабливіших сторін компанії, для залучення клієнтів.
3	Удосконаленя якості послуг	За рахунок отримання параметрів, що найбільше впливають на відтік, може бути прийняте рішення про зміну роботи та модернізацію системи взаємодії з кінцевими користувачами оператора мобільного зв’язку для надання більш якісних послуг ніж у конкурентів.	Застосування новітніх математичних методів, для покращення наданих послуг кінцевим абонентам.

Одночасно і можливістю і загрозою є великі обсяги та складність даних, які отримують оператори зв'язку під час своєї роботи, це дає можливість використовувати новітні методи та алгоритми обробки даних, але одночасно зростає можливість зробити помилку при виявленні параметрів, які найбільше впливають на відтік.

Таблиця 4.8

Ступеневий аналіз конкуренції на ринку

Особливості конкурентного середовища	В чому проявляється дана характеристика	Вплив на діяльність підприємства
1. Тип конкуренції: олігополія	На ринку представлені декілька компаній, що поставляють подібні послуги рішення проблеми відтоку	Акцентування переваг продукту, що забезпечує використання нових методів машинного навчання
2. Рівень конкурентної боротьби: національний/інтернаціональний	Першим етапом є боротьба за ринок України з подальшим виходом на міжнародний ринок.	Маркетингова компанія в першу чергу орієнтована на захоплення місцевого ринку
3. Галузева ознака: внутрішньогалузева	Економічна боротьба з конкурентами відбувається в одній галузі економіки, пропонуються аналогічні послуги, що мають архітектурні відмінності у функціонуванні	Пропозиція суттєвих переваг у порівнянні з продуктами конкурентів у визначеній галузі економіки.

Продовження Таблиці 4.8

Особливості конкурентного середовища	В чому проявляється дана характеристика	Вплив на діяльність підприємства
4. Конкуренція за видами товарів: товарно-видова	Конкуренція відбувається між послугами одного виду. За такої конкуренції значення набуває марка товару.	Постійна робота над забезпеченням високого рівня іміджу компанії
5. За характером конкурентних переваг: нецінова	Передбачається ведення конкурентної боротьби не за рахунок зниження ціни на аналогічні послуги, а за рахунок новизни та унікальних характеристик технології, на якій базується функціонування системи.	Акцент на унікальних характеристиках пропонованого товару.
6. За інтенсивністю: марочна	Виведення товару на ринок передбачається під власною маркою, а також створення асоціації між назвою фірми і методів машинного навчання.	Просування продукту компанії під визначеним брендом.

Таблиця 4.9

Аналіз конкуренції в галузі за М. Портером

Складові аналізу	Прямі конкуренти в галузі	Потенційні конкуренти	Постачальники	Клієнти	Товари-замінники
	SAS, Oracle Big data	Гнучкі ціни, патент на продукт	Змінні витрати постачальників	Рівень чутливості до зміни цін	Ціна, лояльності споживачів
Висновки	Конкуренція не інтенсивна, кожен працює в окремому регіоні.	Можливість входу в ринок висока. Потенційні конкуренти присутні.	Постачальник може диктувати умови: ціни на послуги.	Кожен з клієнтів потребує індивідуального підходу для вирішення його задач	Обмежень для роботи на ринку з боку товарів замінників на даний момент не існує.

В результаті проведення аналізу таблиці 4.9, можна зробити висновок, що можливість виходу на ринок з огляду на конкурентну ситуацію є високою. Для виходу на ринок товар в першу чергу повинен пропонувати унікальні характеристики, які відсутні у продуктах конкурентів.

На основі аналізу конкуренції, проведеного в таблиці 4.9, а також із урахуванням характеристик ідеї проекту (таблиця 4.2), вимог споживачів до товару (таблиця 4.5) та факторів маркетингового середовища (таблиці 4.6 та 4.7), визначається та обґрунтовується перелік факторів конкурентоспроможності, що надається у таблиці 4.10.

Таблиця 4.10

Обґрунтування факторів конкурентоспроможності

№ п/п	Фактори конкурентоспроможності	Обґрунтування
1	Динаміка галузі	Проблема передбачення відтоку абонентів наразі є дуже важливою, тому оператори мобільного зв’язку зацікавлені у системі.
2	Концепція товару і послуги	Система підвищення лояльності клієнтів дозволяє утримувати клієнтів впливаючи на них маркетинговими методами.
3	Після продажне обслуговування	Підтримка щодо використання системи після її продажу.

Таблиця 4.11

Порівняльний аналіз сильних та слабких сторін системи підвищення лояльності абонентів операторів мобільного зв’язку

№ п/п	Фактори конкурентоспроможності	Бали 1-20	Рейтинг товарів-конкурентів у порівнянні з власною системою						
			-3	-2	-1	0	1	2	3
1	Динаміка галузі	3							✓
2	Концепція товару і послуги	1					✓		
3	Після продажне обслуговування	2						✓	

Фінальним етапом ринкового аналізу можливостей впровадження проекту є складання SWOT-аналізу (матриці аналізу сильних (Strength) та слабких (Weak)

сторін, загроз (Troubles) та можливостей (Opportunities) на основі виділених ринкових загроз та можливостей, та сильних і слабких сторін (таблиця 4.12).

Таблиця 4.12

SWOT-аналіз стартап проекту

Сильні сторони: • Інноваційні технології • Висока якість • Комбінація методів	Слабкі сторони: • Слабкий імідж компанії • Слабкий маркетинг • Мало оборотних коштів • Невідома торгівельна марка
Можливості: • Нові технології машинного навчання • Зростання потреб клієнтів • Тенденції попиту	Загрози: • Продукти-замінники

Таблиця 4.13

Альтернативи ринкового впровадження стартап-проекту

№ п/п	Опис профілю цільової групи потенційних клієнтів	Готовність споживачів сприйняти продукт	Орієнтовний попит в межах цільової групи	Інтенсивність конкуренції в сегменті	Простота входу у сегмент
1	Малий бізнес	Переважно готові	Дуже високий	Низька	Легко

Продовження Таблиці 4.13

№ п/п	Опис профілю цільової групи потенційних клієнтів	Готовність споживачів сприйняти продукт	Орієнтовний попит в межах цільової групи	Інтенсивність конкуренції в сегменті	Простота входу у сегмент
2	Всеукраїнські оператори мобільного зв’язку	Не готові	Дуже високий	Висока	Важко
Цільовими групами обрано компанії малого бізнесу, що зацікавлені у зменшення відтоку клієнтів які користуються послугами телекомунікаційної компанії, шляхом застосовування автоматизованих систем.					

Базові стратегії в обраних сегментах ринку представлені у таблиці 4.14.

Таблиця 4.14

Визначення базової стратегії розвитку

№ п/п	Обрана альтернатива розвитку	Стратегія охоплення ринку	Ключові конкурентоспроможні позиції	Базова стратегія розвитку
1	Динамічний розвиток з використанням маркетингу та встановлення бізнес-контактів	Підняття рейтингу компанії шляхом маркетингу, встановлення конкурентоспроможних цін	Незалежність від посередника, який утримує кошти за свої послуги	Стратегія лідерства по витратах

Продовження Таблиці 4.14

№ п/п	Обрана альтернатива розвитку	Стратегія охоплення ринку	Ключові конкурентоспроможні позиції	Базова стратегія розвитку
2	Динамічний розвиток завдяки висвітленню унікальних характеристик надаваємих послуг	Унікальність послуг, для збільшення лояльності клієнта	Використання технології машинного навчання	Стратегія диференціації

Залежно від міри сформованості галузевого ринку, характеру конкурентної боротьби, необхідно обрати одну з трьох стратегій конкурентної поведінки: розширення первинного попиту, оборонну або наступальну стратегію або ж застосувати демаркетинг або диверсифікацію (таблиця 4.15).

Таблиця 4.15

Визначення базової стратегії конкурентної поведінки

№ п/п	Чи є проект першопрохідцем на ринку	Чи буде компанія шукати нових споживачів, або забирати існуючих у конкурентів	Чи буде компанія копіювати основні характеристики	Стратегія конкурентної поведінки
1	Проект не є першопрохідцем	Компанія буде шукати нових користувачів	Компанія буде копіювати найкращі з характеристик конкурентів	Стратегія наслідування лідеру за для економії фінансових ресурсів

На основі вимог споживачів з обраних сегментів до постачальника (стартап-компанії) та до продукту (таблиця 4.5), а також в залежності від обраної базової стратегії розвитку та стратегії конкурентної поведінки була розроблена стратегія позиціонування (таблиця 4.16).

Таблиця 4.16

Визначення стратегії позиціонування

№ п/п	Вимоги до товару цільової аудиторії	Базова стратегія розвитку	Ключові конкурентоспроможні позиції власного стартап-проекту	Вибір асоціацій, які мають сформувати комплексну позицію власного проекту
1	Висока доступність	Стратегія диференціації	Використання методів машинного навчання, що є новим підходом к рішенню цієї задачі.	Доступність, якість, швидкість

4.4 Розроблення маркетингової програми стартап-проекту

Маркетингова програма - система взаємозалежних заходів, що визначають дії підприємства-виробника на заданий період часу з усіх питань маркетингової діяльності. Формування програм маркетингу відбувається на підставі даних щодо комплексного дослідження ринку, визначення поточних і перспективних потреб і попиту потенційних споживачів, з урахуванням обраної стратегії і тактики маркетингу. Програма маркетингу є сполучною ланкою між збутовими і комерційними службами підприємства і науково-технічними, конструкторськими, технологічними та виробничими службами.

Першим кроком є формування маркетингової концепції товару, який отримає споживач. Для цього потрібно підсумувати результати попереднього аналізу конкурентоспроможності товару (таблиця 4.17).

Таблиця 4.17

Визначення ключових переваг концепції потенційного товару

№ п/п	Потреба	Вигода, яку пропонує товар	Ключові переваги перед конкурентом
1	Проведення аналітики великих даних.	Аналіз великих даних забезпечується новими і прогресивними методами машинного навчання.	Гнучкий підхід до рішення поставленної задачі.
2	Отримання переліку закономірностей і факторів, які впливають на відтік абонентів.	Дозволяє розглядати кожен параметр телекомунікаційної мережі окремо ті визначити його вплив.	Подання інформації про параметри у вигляді вибірки для зручності користування.
3	Отримання вибірки абонентів, що мають намір відмовитись від посгул телеком оператора.	Отримання вибірки абонентів з вказанням у відсотках про ймовірність відтоку.	Наочне подання інформації для більш зручного користування.

Надалі розробляється трирівнева маркетингова модель товару: уточняється ідея продукту та послуги, його фізичні складові, особливості процесу його надання (таблиця 4.18).

Таблиця 4.18

Опис трьох рівнів моделі товару

Рівні товару	Сутність та складові
1. Товар за задумом	Товар забезпечує передбачення відтоку абонентів від операторів мобільного зв’язку
2. Товар у реальному виконанні	Властивості: доступність, цілісність, зручність, прозорість, гнучкість.
	Товар представляє собою програмний комплекс виконаний за допомогою мови програмування Python та суміжних бібліотек.
	Поставляється у вигляді застосунку для всіх популярних платформ.
	Назва: Predicted of customer churn in telecommunication network
3. Товар із підкріпленням	До продажу: відбувається інсталяція та конфігурування системи, проводяться тренінги по користуванню для клієнта
	Після продажу: відбувається підтримка програмного забезпечення та його доопрацювання під потреби клієнта
Програмний комплекс, що забезпечує передбачення відтоку та підвищення лояльності абонентів, розповсюджується на основі підписки, захисту підлягає програмний код та програмна реалізація.	

Аналіз системи збуту передбачає визначення ефективності кожного елемента цієї системи, оцінювання діяльності апарату працівників збуту. Аналіз витрат обігу передбачає зіставлення фактичних збутових витрат за кожним каналом збуту і видом витрат із запланованими показниками для того, щоб виявити необґрунтовані витрати, ліквідувати затрати, що виникають у процесі руху товарів і підвищити рентабельність наявної системи збуту.

Дані щодо визначення системи збуту надаються в таблиці 4.19.

Таблиця 4.19

Формування системи збуту

№ п/п	Специфіка закупівельної поведінки цільових клієнтів	Функції збуту, які має виконувати постачальник товару	Глибина каналу збуту	Оптимальна система збуту
1	Власна система збуту	Проведення та розгортання програмного забезпечення на стороні клієнта (телеком оператора)	Канал нульового рівня, продаж товару відбувається безпосередньо споживачам через відділ збуту.	Оптимальною системою збуту є прямий збут з каналом нульового рівня за відсутності посередників

Таблиця 4.20

Концепція маркетингових комунікацій

№ п/п	Специфіка поведінки цільових клієнтів	Канали комунікацій	Ключові позиції, обрані для позиціонування	Завдання рекламного повідомлення
1	Консервативна поведінка, але відкриті до нового.	Соцмережі професійного спрямування, корпоративна пошта	Можливості отримання переліку параметрів та абонентів	Висвітлити унікальні характеристики продукту

У якості концепції маркетингових комунікацій були обрані інтегровані маркетингові комунікації, де компанія ретельно обмірковує і координує роботу своїх численних каналів комунікації, рекламу в засобах масової інформації, особистий продаж, стимулювання збуту, пропаганду, прямий маркетинг.

Висновки

В даному розділі був проведений маркетинговий аналіз перспектив реалізації системи визначення лояльності абонентів для телеком операторів та проведене оцінювання можливостей її ринкового впровадження.

В результаті дослідження визначено, що існує можливість ринкової комерціалізації проекту в першу чергу завдяки використанню технологій машинного навчання, що дозволяє наділити продукт унікальними

характеристиками, такими як швидкість, прозорість, гнучкість, якість отриманого рішення та великі можливості кастомізації проекту.

Конкурентна ситуація надає перспективи впровадження продукту, так як продукція товарів-аналогів має лише частковий функціонал реалізованої системи та володіє низкою критичних недоліків, через які рівень довіри до них залишається незадовільним. В результаті існуючі товари-аналоги не створюють прямої конкуренції на ринку України. Основною проблемою є можливе негативне ставлення всеукраїнських операторів мобільного через незнання переваг нашої системи.

Проведений аналіз підтверджує, що подальша імплементація проекту є доцільною.

ЗАГАЛЬНІ ВИСНОВКИ ПО РОБОТІ

Магістерська робота становить собою закінчене наукове дослідження, в якому розв’язано актуальну науково-технічну задачу підвищення ефективності виявлення факторів та закономірностей із статистичних наборів даних, які впливають на лояльніть абонента.

Отримані наукові результати, що мають істотні переваги перед існуючими рішеннями:

1. Проведено аналіз методів машинного навчання та бізнес процесів, які застосовуються для аналізу статистичних даних в телекомунікаційних системах, який показав, що існує проблема підвищення якості оперування, управління та обробки великих об’ємів даних, спричинена її слабкою структурованістю, недостатньою систематизованістю, різнорідністю та слабкозв’язаністю. Оскільки вся інформація, яка збирається під час роботи телекомунікаційної мережі, є неоднорідною, то для обробки необхідно використовувати комбінацію з різноманітних методів.
2. Проведено аналіз вхідних даних, які отримані під час роботи однієї з великих українських операторів мобільного зв’язку.
3. Було поведено математичне моделювання методів передбачення відтоку клієнтів за допомогою методів асоціативних правил, дерева рішень та bagging виявлено, що метод bagging дозволяє отримати більш точні результати, на 7% в порівнянні з методом асоціативних правил та та до 3% кращі ніж значення метрики методу дерева рішень, точність отриманих результатів складає 80,45%.
4. Визначено закономірності та фактори, які найбільш впливають на рішення абонента відмовитись від послуг оператора зв’язку, а саме:
    - тривалість користування послугами телекомунікаційної компанії;

- кількість днів, під час якого абонент не користується послугами;
- загальна кількість дзвінків інших мобільних операторів;
- вихідні дзвінки на номери інших мобільних операторів;
- загальна кількість міжнародних дзвінків.

5. В результаті проведення досліджень виявлено вибірку абонентів, які схильні до відтоку та складена класифікація абонентів в залежності від відсотку лояльності.
6. Запропоновано сценарій бізнес процесу для вирішення проблеми передбачення відтоку абонентів на основі комбінації методів машинного навчання для покращення якості наданих послуг.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. «Онемели: кто из мобильных операторов заработал больше всего,» delo.ua, 10 грудня 2018. https://delo.ua/business/onemeli-kto-iz-mobilnyh-operatorov-zarabotal-bolshe-vsego-337433

2. В. Скорбота, «Как украинцы выбирают мобильного оператора,» 28 вересня 2017. https://biz.nv.ua/experts/skorbota/kak-ukraintsy-vybirajut-mobilnogo-operatora-1930612.html.

3. J. Han, M. Kamber та J. Pei, Data Mining: Concepts and Techniques, Elsevier, 2011.

4. A. Ng, «CS229 Lecture notes,» https://see.stanford.edu/materials/aimlcs229/cs229-notes1.pdf.

5. L. Breiman, «Random Forests,» *Machine Learning*, № 45, p. 5–32, October 2001.

6. L. Breiman, «Bagging Predictors,» в *Machine Learning*, 1996.

7. Y. Sasaki, «The truth of the F-measure,» 26th October 2007. https://www.toyota-ti.ac.jp/Lab/Denshi/COIN/people/yutaka.sasaki/F-measure-YS-26Oct07.pdf.

8. «Анализ с помощью характеристической кривой,» [Онлайновий]. Available: https://wiki.loginom.ru/articles/roc-analysis.html.

9. «About Python,» https://www.python.org/about/.

10. «About Python,» https://www.python.org/2.5/license.html.

11. R. Cotton, Learning R, Sebastopol, CA: O'Reilly Media, 2013.

12. А. Е. Курбатова, MATLAB 7. Самоучитель, Москва: «Диалектика», 2005.

13. П. В. Дьяконов, Справочник по применению системы PC MATLAB, Москва: «Физматлит», 1993.

14. «Аналитика, бизнес-аналитика и управление данными | SAS,» 2018. https://www.sas.com/ru_ru/home.html.

15. «Все, что нужно знать о мошенничестве в машинном обучении | SAS,» SAS, 2017. https://www.sas.com/ru_ua/insights/articles/analytics/fraud-detection-machine-learning.html#/.

16. «Python Documentation — Built-in Types,» https://docs.python.org/3/library/stdtypes.html.

17. «Are tuples more efficient than lists in Python? — Stack Overflow,» https://stackoverflow.com/questions/68630/are-tuples-more-efficient-than-lists-in-python.

18. «eGenix.com — Professional Python Software, Skills and Services,» http://www.egenix.com/.

19. «numarray Home Page,» http://www.stsci.edu/institute/software_hardware/numarray.

20. «PEP333,» https://www.python.org/dev/peps/pep-0333/.

21. «Pyste Documentation,» http://www.boost.org/doc/libs/1_64_0/libs/python/pyste/index.html.

22. «SIP,» https://riverbankcomputing.com/sip/.

23. «Pyfort,» http://pyfortran.sourceforge.net/.

24. «Boost.Python,» http://www.boost.org/doc/libs/1_64_0/libs/python/doc/html/index.html.

25. «Building Hybrid Systems with Boost.Python,» http://www.drdobbs.com/building-hybrid-systems-with-boostpython/184401666.

26. «PyCXX: Write Python Extensions in C,» http://cxx.sourceforge.net/.

27. «Мост между C++ и Python http://pyhrol.ru/wiki/Pyhrol.html.

28. «PyInline: Mix Other Languages directly Inline with your Python,» http://pyinline.sourceforge.net/.

29. «Weave,» https://www.scipy.org/Weave.

30. Ш. Франсуа, Глубокое обучение на Python, Санкт-Петербург: "Издательский дом ""Питер", 2018.

31. Ф. Шолле, Глубокое обучение на R, Санкт-Петербург: "Издательский дом ""Питер", 2018.

32. Пальмов С. В. Порівняння можливостей різні методів технології Data Mining при аналізі персонального трафіку. // XII Российска наукова конференція професорсько-викладацького складу, наукових співробітників та аспірантів, Самара, ПДАТУ, 2005, теза доповідей, стор. 285-287.

33. Барсегян А.А., Купріянов М.С. і ін. Методи і моделі аналізу даних: OLAP і Data Mining. - СПб .: БХВ-Петербург, 2004

34. «Онемели: кто из мобильных операторов заработал больше всего,» delo.ua, 10 грудня 2018. [Онлайновий]. Available: https://delo.ua/business/onemeli-kto-iz-mobilnyh-operatorov-zarabotal-bolshe-vsego-337433.

35. В. Скорбота, «Как украинцы выбирают мобильного оператора,» 28 вересня 2017. [Онлайновий]. Available: https://biz.nv.ua/experts/skorbota/kak-ukraintsy-vybirajut-mobilnogo-operatora-1930612.html.

36. Виявлення знань та досвіду (емпіричних фактів) і інтелектуальний аналіз даних (data mining) www.Ic.kubargo.ru/aidos/ASK-Analis/LK-14/lk-14.htm#_Toc78426168

37. The truth of the F-measure Y. Sasaki, Version: 26th October, 2007

38. «About Python,» [Онлайновий]. Available: https://www.python.org/about/

39. L. Breiman, «Random Forests,» Machine Learning, № 45, p. 5–32, October 2001.

40. «Анализ с помощью характеристической кривой,» [Онлайновий]. Available: https://wiki.loginom.ru/articles/roc-analysis.html.

41. «About Python,» [Онлайновий]. Available: https://www.python.org/2.5/license.html.

42. П. В. Дьяконов, Справочник по применению системы PC MATLAB, Москва: «Физматлит», 1993.

43. Ф. Шолле, Глубокое обучение на R, Санкт-Петербург: "Издательский дом ""Питер", 2018.

44. І.Л. Кафтанніков, А.В. Парасич, Особливості застосування дерев рішень в задачах Південно-Уральський державний університет, м Челябінськ.

45. Д. В. Нечипорук, Особливості технології Data Mining, Донський державний технічний університет, Ростов-на-Дону, Російська Федерація.

46. Дерікьянц Павло Павлович, Огляд архітектурі сучасних білінгових систем і перспективи їх розвитку.

47. A.L. Samuel Some Studies in Machine Learning Using the Game of Checkers// IBM Journal. July 1959. P. 210–229.

48. Айвазян С.А., Енюков І.С., Мешалкин Л.Д. Прикладна статистика: основи моделювання і первинна обробка даних. М .: Фінанси і статистика, 1983. 471 с.

49. Самсонов М. Тенденції ринку телекомунікаційного ПО // ІнформКурьер-Зв'язок. 2003. № 2.

50. АЛГОРИТМИ КЛАСТЕРИЗАЦІЇ В ЗАДАЧАХ СЕГМЕНТАЦІЇ СУПУТНИКОВИХ ЗОБРАЖЕНЬ, І. А. Пестунов, Ю. Н. Синявський

51. Аналіз і класифікація алгоритмів кластеризації, Єршов К.С., Романова Т.Н., МГТУ ім. Н.Е. Баумана

52. A Simple Approach to Predicting Customer Churn, Leo YorkeLewis Fogden, Thu 29 June 2017, in category Data science, http://blog.keyrus.co.uk/a_simple_approach_to_predicting_customer_churn.html

53. Big Data (Великі дані), онлайн ресурс https://www.it.ua/knowledge-base/technology-innovation/big-data-bolshie-dannye

54. Отаманів, Ю. С. Введення в Big Data / Ю. С. отаманів, В. С. Гончарук, С. Н. Гордєєв. - Молодий вчений. - 2017. - № 11 (145). - С. 33-34. - URL: https://moluch.ru/archive/145/40562/

55. Бакалаврська робота «Аналіз та прогнозування відтоку клієнтів в телекомунікаційній компанії на основі технології Data Mining» - Мороз А.М., 2019

56. Студентська наукова робота «Аналіз та прогнозування лояльності абонентів в телекомунікаційній мережі на основі технології машинного навчання» Мороз А.М., 2020

57. Методичні рекомендації до виконання розділу магістерських дисертацій для студентів інженерних спеціальностей «РОЗРОБЛЕННЯ СТАРТАП-ПРОЕКТУ», Київ НТУУ «КПІ ім. Ігоря Сікорського», 2016

ДОДАТОК А

Таблиця 1

Найменування параметрів

#	Назва	Опис	Діапазон значень
1	ABON_CODE	Хеш код абонента	1-19999
2	STATUS	статус абонента	0; 1
3	COUNT_DAYS_OVER_1MB	Кількість днів протягом місяця більше 1 Мб	1-31
4	COUNT_DAYS_OVER_5MB	Кількість днів протягом місяця більше 5 Мб	1-31
5	DUAL_SIM_PROBABILITY	Імовірність мати 2 сім карти	Very low; Low; High; Very high; UNDEF
6	SIM_PRIORITY	Пріоритетність сімки	FIRST; SECOND; UNDEF
7	OBLAST	Область	
8	CITY	Місто	
9	CN_ONNET_6M_TILE4	кількість оператору 1	1-4
10	CN_ONNET_SHARE_6M_TILE4	кількість оператору 1	1-4
11	CN_ONNET_6M	Загальна кількість дзвінків оператору 1	0-7101
12	CN_OMO_6M	Загальна кількість дзвінків інших мобільних операторів	0-11334
13	CN_INTL_6M	Загальна кількість міжнародних дзвінків	0-170
14	CN_ONNET_SHARE_6M	кількість оператору 1	0-1.0
15	N_SERVICES	Кількість діючих сервісів	1-7
16	ONNET_DIFF_B_OUT_6M	Вихідні дзвінки на оператору 1	1-7101

17	OP2_DIFF_B_OUT_6M	Вихідні дзвінки на оператор 2	1-11334
18	OP3_DIFF_B_OUT_6M	Вихідні дзвінки на оператор3	1-2261
19	NTL_OMO_DIFF_B_OUT_6M	Вихідні дзвінки на номери інших мобільних операторів	1-37
20	NTL_PSTN_DIFF_B_OUT_6M	Вихідні дзвінки на номери Укртелеком	1-1573
21	RUS_MO_DIFF_B_OUT_6M	Вихідні дзвінки в Росію на мобільні	1-18
22	RUS_PSTN_DIFF_B_OUT_6M	Вихідні дзвінки в Росію на міські	1-16
23	OTHER_INTL_DIFF_B_OUT_6M	Вихідні дзвінки на міжнародні номери	1-169
24	ONNET_DIFF_A_INC_6M	Вхідні дзвінки з оператору 1	1-3622
25	OP2_DIFF_A_INC_6M	Вхідні дзвінки з оператору 2	1-9141
26	OP3_DIFF_A_INC_6M	Вхідні дзвінки з оператору 3	1-2262
27	NTL_OMO_DIFF_A_INC_6M	Вхідні дзвінки з інших мобільних операторів	1-55
28	NTL_PSTN_DIFF_A_INC_6M	Вхідні дзвінки з номерів Укртелекому	1-544
29	RUS_MO_DIFF_A_INC_6M	Вхідні дзвінки з Росії з мобільних номерів	1-26
30	RUS_PSTN_DIFF_A_INC_6M	Вхідні дзвінки з Росії з міських номерів	1-8
31	OTHER_INTL_DIFF_A_INC_6M	Вхідні дзвінки з міжнародних номерів	1-167
32	TENURE	Тривалість користування послугами мобільного оператора	1-141

33	TARIFF_CHANGE	поповнення	0-1
34	AVG_DAYS_BETW_REF_6M	Середнє значення днів перед поповненням	0-180.0989
35	REFILL_FREQ_6M	Поповнення кожної неділі	0.1667-24.1667
36	AVG_REFILL_AMOUNT_6M	Середнє значення поповнень в місяць	0.49-1000
37	AVG_BALANCE_BEFORE_REF_6M	Середнє значення балансу перед поповненням	(-36.4862)-7663.1253
38	SILENT_MONTHS_12	Кількість місяців бездіяльності	0-12
39	DAYS_INACT_ALL_6	Кількість не активних днів	0-184
40	AVG_DAYS_ACT_ALL_6	Середнє значення всіх активних днів	0-31
41	AVG_DAYS_INACT_ALL_6	Середнє значення всіх не активних днів	0-30.67
42	AVG_DAYS_ACT_SUCC_6	Середнє значення активних днів	0-31
43	AVG_DAYS_INACT_SUCC_6	Середнє значення не активні днів	0-30.67
44	TOP_CONTACTS_MTC	Топ 5 контактів пішли за останній звітний місяць	0-32
45	TOP_CONTACTS_ALL	Топ 5 контактів пішли за весь час	0-26
46	AVG_ONNET_OUT_COUNT_CALLS_6M	Середнє значення довжини вихідних дзвінків на оператора 1	0-2061.1667
47	AVG_OMO_OUT_COUNT_CALLS_6M	Середнє значення довжини вихідних дзвінків на номери інших мобільних операторів	0-776.5

48	AVG_ONNET_INC_COUNT_CALLS_6M	Середнє значення довжини вхідних викликів з оператора 1	0-2974.3333
49	AVG_OMO_INC_COUNT_CALLS_6M	Середнє значення довжини дзвінки що надходять з інших мобільних операторів	0-1808.5
50	MINS_SLOPE	Тангенс кута нахилу лінійного тренда для кількості хвилин потижнево	(-0.999999908256889)-0.999999845916819
51	INET_SLOPE	Тангенс кута нахилу лінійного тренда для обсягу даних потижнево	(-0.999999999999958)-0.999999999999924
52	REFILL_SLOPE	Тангенс кута нахилу лінійного тренда для поповнення потижнево	(-0.999999866666684)-0.999999896907227
53	INS_STD	СКО	0-0.707106732081913
54	INET_STD	СКО (середньоквадратичне відхилення) для обсягу даних потижнево	0-0.707106781186547
55	REFILL_STD	СКО для поповнення потижнево	0-0.707106730135596
56	MINS_REG_CONST	«Підйом» лінійного тренда для кількості хвилин потижнево	(-201646.968929589)-201644.49588876
57	INET_REG_CONST	«Підйом» лінійного тренда для	(-201646.999990154)-201646.999999987
58	REFILL_REG_CONST	обсягу даних потижнево	(-201646.938894867)-201647.844886273
59	MINS_SLOPE2	«Підйом» лінійного тренда для поповнення потижнево	(-0.999999996845426)-0.999999968457543

60	INET_SLOPE2	Тангенс кута нахилу лінійного тренда для кількості хвилин потижнево за 2 період часу	(-0.999999999999998)-0.999999999999994
61	REFILL_SLOPE2	Тангенс кута нахилу лінійного тренда для обсягу даних потижнево за 2 період часу	(-0.99999993610224)-0.999999910714294
62	INS_STD2	Тангенс кута нахилу лінійного тренда для поповнення потижнево за 2 період часу	0-0.707106758882663
63	INET_STD2	СКО за 2 період часу	0-0.707106781186546
64	REFILL_STD2	СКО для обсягу даних потижнево за 2 період часу	0-0.707106750173094
65	MINS_REG_CONST2	СКО для поповнення потижнево за 2 період часу	(-201644.887975062)-201643.95709703
66	INET_REG_CONST2	«Підйом» лінійного тренда для кількості хвилин потижнево за 2 період часу	(-201646.999999985)-201647.999999992
67	REFILL_REG_CONST2	«Підйом» лінійного тренда для	(-201646.979835302)-201647.844886273
68	MINS_SLOPE3	Обсяг даних потижнево за 2 період часу	(-0.999999954569617)-0.99999949748769
69	INET_SLOPE3	«Підйом» лінійного тренда для поповнення потижнево за 2 період часу	(-0.999999999999568)-0.999999999999924
70	REFILL_SLOPE3	Тангенс кута нахилу лінійного тренда для кількості хвилин потижнево за 3 період часу	(-0.999999966357154)-0.999999947368424
71	INS_STD3	Тангенс кута нахилу лінійного тренда для обсягу даних потижнево за 3 період часу	0-0.707106752558744

72	INET_STD3	Тангенс кута нахилу лінійного тренда для поповнення потижнево за 3 період часу	0-1670209922369810
73	REFILL_STD3	СКО за 3 період часу	0-21693080.5463342
74	MINS_REG_CONST3	СКО для обсягу даних потижнево за 3 період часу	(-201641.871155354)-201642.990839281
75	INET_REG_CONST3	СКО для поповнення потижнево за 3 період часу	(-201646.999999988)-201647.999999992
76	REFILL_REG_CONST3	«Підйом» лінійного тренда для кількості хвилин потижнево за 3 період часу	(-201646.98418455)-201647.953105127